Human/Machine Interface Modalities for Soldier Systems Technologies

20030617 134

REPORT DOCUMENTATION PAGE

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE October 25, 2002	3. REPORT TYPE AND DATES COVERED Final Report (4/1/2002 – 10/31/2002)	

4. TITLE AND SUBTITLE Human/Machine Interface Modalities for Soldier Systems Technologies	5. FUNDING NUMBERS C/TA
6. AUTHOR(S) Sandeep Mulgund, John Stokes, Melanie Turieo, and Marlene Devine	

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) TIAX, LLC Acorn Park Cambridge, Massachusetts 02140	8. PERFORMING ORGANIZATION REPORT NUMBER 71950-00

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Natick Soldier Center SBCCOM ATTN: Cynthia Blackwell Natick, Massachusetts 01760	10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution is unlimited	12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

The Army's Objective Force Warrior program seeks to create a lightweight, overwhelmingly lethal, fully integrated individual combat system. This includes weapon, head-to-toe individual protection, networked communications, soldier-worn power sources, and enhanced human performance. Achieving this objective will in part entail the development of soldier-centric human/machine interfaces (HMIs) that optimize cognitive fightability. Such optimization is possible only if these HMIs are designed in such a way that takes into account the nature of human information processing and cognition. This in turn depends on understanding how best to use the senses by which humans perceive their environment and the means by which they can affect it; i.e., the *modalities* for human/machine interaction. Traditional approaches to HMI design have centered on the use of visual displays and manual inputs, but these do not take advantage of the full range of means by which humans can perceive and interact with their environment.

This report reviews the literature on human/machine interface modalities. It also provides guidelines for system designers to consider when choosing which modalities should be considered in a system intended to augment human cognitive performance.

14. SUBJECT TERMS human/machine interfaces, multimodal, wearable systems, human factors, objective force warrior	15. NUMBER OF PAGES 70
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	

Executive Summary

The U.S. Army's Objective Force Warrior program seeks to create a lightweight, overwhelmingly lethal, fully integrated individual combat system. This includes weapon, head-to-toe individual protection, networked communications, soldier-worn power sources, and enhanced human performance. Achieving this objective will in part entail the design and development of soldier-centric human/machine interfaces (HMIs) that optimize cognitive fightability. Such optimization is possible only if these HMIs are designed in such a way that takes into account the nature of human information processing and cognition. This in turn depends on understanding how best to use the senses by which humans perceive their environment and the means by which they can affect it; i.e., the *modalities* for human/machine interaction. Traditional approaches to HMI design have centered on the use of visual displays and manual inputs, but these do not take advantage of the full range of means by which humans can perceive and interact with their environment.

This report reviews the literature on human information processing as it relates to the selection of HMI modalities. By reviewing the different modalities that can be used for information presentation and system control, it provides guidelines for system designers to consider when choosing which modalities should be considered in a system intended to augment human cognitive performance. Four key areas are covered in this report:

- Models of Human Information Processing
- Modalities for Information Presentation
- Modalities for System Control
- Analysis of Soldier Needs

Each is summarized below.

Models of Human Information Processing

Research in the area of cognitive psychology provides a conceptual framework for system designers to make sensible choices about how HMIs should be designed to optimize human performance. In particular, the following considerations motivate an appropriate high-level view of task-specific HMI design:

- A classification of **human behavior** based on the degree of cognitive processing
- **Compatibility in stimulus-central processing-response** for display/control design
- Effects of **attention-sharing** and **multi-tasking**

Human behavior can be classified into three successive levels: skill-based, rule-based, and knowledge-based. Each has a higher degree of cognitive processing and a lesser amount of automaticity than the last. Effective execution of skill-based actions depend to a large degree on the *speed* of human response, while *accuracy* is often most important for rule-based or knowledge-based activities. When choosing modalities for information presentation, consideration should therefore be given to whether the underlying task is inherently rule- or knowledge-based (placing the premium on

accuracy and precision of information presentation), or skill-based (placing the premium on speed and response time). Simple metrics of reaction time are typically optimized through the use of auditory alerting, but visual presentation is preferable when accuracy is important or continuous feedback is required (e.g., for manual tracking tasks).

A key consideration in modality selection is the notion of the compatibility of **stimulus-response** (SR) pairing in a display/control relation. The literature indicates that visual inputs are well-paired with manual outputs and auditory inputs with speech output, when the performance criterion is a measurement of reaction time. While this provides a starting point, it does not consider the requirements of rule-based or knowledge-based tasks, nor does it account for multi-tasking effects. This notion of SR compatibility has thus been extended to consider the type of central cognitive processing associated with a task, and generalized as stimulus-central-processing-response (SCR) compatibility. Spatial tasks are those that involve a judgment concerning the axes of translation and rotation, while verbal tasks are those involving the use of language and arbitrary symbolic coding. Spatial tasks tend to be best associated with visual inputs and manual outputs, while verbal tasks match well with auditory input and speech output. A variety of considerations constrain these findings, which are discussed in detail. It is often the case that augmenting primary modalities with others can be very effective.

Finally, attention should be given to the effects of **attention-sharing and multi-tasking**. The *multiple resource theory of human information processing* proposes that the resources for which multiple tasks compete may be defined by input and output modalities, by stages of information processing, and by the codes (verbal/spatial) of processing. These dimensions are thought to be somewhat independent of one another. Consequently, two tasks that share common resource demands will be time-shared less effectively than those with non-overlapping demands.

Modalities of Information Presentation

In any human-machine system, efficient and effective operation relies upon constant exchange of information between the machine and the human user/operator. One direction of flow is from the machine to the human. As there are five senses that allow humans to perceive the world around them, there are in theory five modalities by which information can be presented: visual, auditory, haptic, olfactory, and taste. Most of the literature focuses on the first three. There has been some work in the area of olfactory cueing, but it is not yet a practical means of information cueing. The use of taste for information cueing is largely unexplored. There are therefore three practical modalities available to system designers for information presentation. While in principle any of them may be used in an HMI, each has strengths and weaknesses in conveying different types of information. The applicability and some of the means of exploitation of each of the output modalities is summarized in the table below.

Modality	Methodology	Applicability and Guidelines for Use
Visual	Head-mounted displays	• Overall, visual presentation is applicable for spatial information • Visual textual displays acceptable for verbal information, particularly longer messages • HMDs particularly useful in situations where information is needed without diverting gaze, "on-the-move" operations
	Text vs. Graphic Presentation	• Graphic/symbolic presentation preferred for speed • Textual presentation preferred for accuracy • Can be combined to facilitate speed of comprehension, accuracy of interpretation, and long-term retention
Auditory	Alerts and warnings	• Effective for rapid cueing of critical information • Alerts should be short and simple • Bi-modal (visual with auditory) warnings can elicit faster response than either uni-modal warning
	Spatial Audio Cues	• Audio signals can be spatialized to indicate direction and location and movement • Spatialized audio can help identify auditory messages in noisy conditions or could be used for navigation tasks (waypoints, object/person locations, etc)
Haptic	Tactile Cues	• Effective for simple alerting via vibrations, pressure, etc. • More complex applications to provide sense of presence, orientation, or direction in a task environment without added cognitive burden
	Kinesthetic Cues	• Can supplement other displays to help remember location of items in space relative to self, increasing recall • Increase usability and reduce mode error by implementing quasi-modes that are maintained kinesthetically • Typically not used alone

Modalities of System Control

In addition to the modalities for presenting information to a human user, we must also consider the means by which humans provide input or control commands into a system. In recent years there has been an explosion in the number and capability of devices to facilitate interaction between humans and machines. Ideally, to make this interaction be as natural as possible, computers should be able to interpret all natural human actions. The different control modalities that can be used for providing input to a system via an HMI are: manual, gaze, speech, and neural. Each one has strengths and weaknesses with respect to different types of tasks. While manual control technologies are the most mature, gaze and speech-based control are becoming increasingly viable. Neural control is a relatively new technique that offers some promise. It may prove to be an effective means by which disabled or paralyzed individuals can interact with the technology around them. The table below outlines the applicability and benefits of the modalities that can be used for system control.

Modality	Technology/Methodology	Applicability and Guidelines for Use
Manual	Simple	• Generally suited for controlling spatial elements; often paired with visual presentations • Also used for communicating verbal information such as text entry, but not always ideal • Visual, auditory, or haptic feedback should be provided to indicate status
	Gesture	• Can be used for manipulation and communication
Gaze-Based Control	Head	• Best for infrequent, non-command applications, because long or repeated movements of head can become annoying, tiresome, and distracting from other tasks • Best used for tasks where the control is closely associated with gaze such as moving or selecting items • Can be considered in circumstances where physical impairment limits usage of other control modalities
	Eye	• More applicable for frequently-used controls because the eye can better handle frequent, repetitive movements
Speech Recognition	Speech-Based Browsing	• Verbal commands and text entry can be "more natural" with speech compared to typing or handwriting recognition • Allows user to activate functions without diverting gaze
	Command and Control	• Can be used for just about any operating commands where background noise is not a cause for error • Frees hands to be used in other concurrent tasks

Analysis of Soldier Needs

Information display and controls components associated with soldier systems are typically developed according to the way the information has historically been provided, or on the type of technology that is currently available. Developing information display systems for soldiers by first examining what their information needs are and then determining what the best modality or combination of modalities would be to present that information, would lead to more robust, usable, and effective soldier systems. The type of information being displayed, the action being performed, and the information processing to be done, should drive the development of information display technologies for soldier systems, not vice versa.

In the final chapter of this report we examine how the principles and design methodologies discussed above can be used to support task-oriented identification of HMI modalities for soldier systems. Relevant soldier needs are first identified, based on work conducted under the Scorpion and OFW programs to characterize soldier tasks and activities. Each of the associated tasks are then classified according to the type of cognitive processing they entail. Finally, we present some preliminary thoughts on which HMI modalities offer the potential to best support human performance in supporting these tasks. This analysis is based on making a connection between the nature of the cognitive processing associated with a soldier task and the findings from the literature on the merits and applicability of various input/output modalities. The selection of modalities is meant only to illustrate how the principles discussed in this report can be used to motivate human-centric systems design. Exact specification requires detailed consideration of tasks, context, and concurrent activity.

Table of Contents

List of Tables

List of Figures

1. Overview

1.1 Introduction

The U.S. Army's Objective Force Warrior program seeks to create a lightweight, overwhelmingly lethal, fully integrated individual combat system. This includes weapon, head-to-toe individual protection, networked communications, soldier-worn power sources, and enhanced human performance. Achieving this objective will in part entail the design and development of soldier-centric human/machine interfaces (HMIs) that optimize cognitive fightability. Such optimization is possible only if these HMIs are designed in such a way that takes into account the nature of human information processing and cognition. This in turn depends on understanding how best to use the senses by which humans perceive their environment and the means by which they can affect it; i.e., the *modalities* for human/machine interaction. Traditional approaches to HMI design have centered on the use of visual displays and manual inputs, but these do not take advantage of the full range of means by which humans can perceive and interact with their environment.

This report reviews the literature on human information processing as it relates to the selection of HMI modalities. By reviewing the different modalities that can be used for information presentation and system control, it provides guidelines for system designers to consider when choosing which modalities should be considered in a system intended to augment human cognitive performance. While these guidelines can never be as immutable as physical laws of nature, there is a sufficiently large body of research in the literature to justify the rational selection of alternate modalities that can improve upon the traditional visual/manual combination.

1.2 Overview of Report

This report contains four additional chapters and a list of references.

Chapter 2 provides an overview of relevant theories and models of human information processing that relate to the rational selection of human/machine interface modalities. These models help to provide a theoretical and empirical framework that allows system designers to make sensible choices about how HMIs should be designed to optimize human performance. Section 2.2 presents a taxonomy of skilled human behavior, categorizing human behavior based on its degree of *automaticity*. This automaticity often relates to the amount of practice that a human has in application of a certain skilled-based behavior, and has consequences for the design of systems that support those tasks. In section 2.3 we present an overview of research on the subject of stimulus/response compatibility. The results of this research help to identify what types of human processing and response tasks are suited to a given modality of input and response. Section 2.4 discusses the issues associated with attention sharing and multi-tasking. Finally, section 2.5 summarizes the chapter's contents.

Chapter 3 discusses the modalities and technologies used for presenting information to the human user. As there are five senses that allow humans to perceive the world around them, there are in theory five basic modalities by which information can be presented in

a human/machine interface: visual, auditory, haptic/tactile, olfactory, and taste. Most of the research and available literature focuses on the first three – visual, auditory, and tactile. These modalities are covered in sections 3.2 to 3.4, respectively. There has been some work in the area of olfactory interfaces, which is discussed in section 3.5. However, it is not an area of any significant active research, and we do not anticipate an application for this approach in the foreseeable future. The use of taste as a stimulus mechanism in human/machine interfaces is largely unexplored, and not addressed in this report. Section 3.6 summarizes chapter's contents.

Chapter 4 provides an overview of the modalities and technologies that may be used to enable a human user to control a computer-based system. Section 4.2 reviews manual control, which is the more traditional and most exploited modality for user input. Gesture-based controls are discussed as a subset of manual control. Section 4.3 considers both head and eye gaze-based controls. Section 4.4 reviews speech interfaces and various issues surrounding the use of speech-based interactions. Section 4.5 reviews recent explorations in neural control. Each of these sections deals with the modality primarily by itself. Section 4.6 then explains multimodal control systems and the rationale for using them. Finally, section 4.7 summarizes the chapter's contents.

Finally, chapter 5 connects the discussion in chapters 2-4 with the anticipated needs of the OFW soldier. In this chapter we examine how the principles and design methodologies discussed in the preceding three chapters can be used to support task-oriented identification of HMI modalities for soldier systems. Section 5.2 identifies the relevant soldier needs, based on work conducted under the Scorpion and OFW programs to characterize soldier tasks and activities. In section 5.3 we describe the process by which these tasks were classified according to the type of cognitive processing they entail (following the principles discussed earlier in Chapter 2). Finally, section 5.4 presents some preliminary thoughts on which HMI modalities offer the potential to best support human performance in supporting these tasks. This analysis is based on making a connection between the nature of the cognitive processing associated with a task and the findings from the literature on the merits and applicability of various input/output modalities, as discussed earlier in chapters 3 and 4.

1.3   Abbreviations and Acronyms

ASL	American Sign Language
CAD	Computer-aided design
EEG	Electroencephalographic
EMG	Electromyographic
EOG	Electrooculographic
HCI	Human/computer interaction
HMD	Head-mounted display
HMI	Human/machine interface
HUD	Heads-up display
MAMA	Minimal audible movement angle
OFW	Objective Force Warrior

PC	Personal computer
RT	Reaction time
SA	Situation awareness
SCR	Stimulus-central processing-response
TSAS	Tactical Situation Awareness System

2. Theories of Human Information Processing

2.1 Introduction

Next-generation soldier systems technologies must be designed to work *cooperatively* with the soldier. If new systems are introduced in an uncoordinated fashion without concern for human cognitive capabilities and limitations, the effect will likely be to confuse, distract, and overwhelm the soldier rather than enhance performance. Ensuring the latter requires an understanding of human information processing and skilled human behavior. In this chapter we provide an overview of relevant theories and models of human information processing that relate to the rational selection of human/machine interface modalities. These models help to provide a theoretical and empirical framework that allows system designers to make sensible choices about how HMIs should be designed to optimize human performance.

Section 2.2 presents a taxonomy of skilled human behavior. This taxonomy categorizes human behavior based on its degree of *automaticity*. This automaticity often relates to the amount of practice that a human has in application of a certain skilled-based behavior, and has consequences for the design of systems that support those tasks. In section 2.3 we present an overview of research on the subject of stimulus/response compatibility. These results help to identify what types of human processing and response tasks are suited to a given modality of input and response. Section 2.4 discusses the issues associated with attention-sharing and multi-tasking. Finally, section 2.5 summarizes the chapter's contents.

2.2 A Taxonomy of Skilled Human Behavior

Figure 2-1 below presents Rasmussen's hierarchy of skilled human behavior, which categorizes human behavior according to the degree of automatic processing (Rasmussen, 1980, 1986). He describes human behavior in three successive levels:

- **Skill-based** behavior, which is continuous, effectively automatic, well-learned sensorimotor behavior. Stimuli are assigned to responses in a rapid automatic fashion with a minimal investment of cognitive resources.
- **Rule-based** behavior, which relates to recognizing pattern of stimuli and triggering "if-then" algorithms to execute the appropriate response. This behavior is demonstrated in situations requiring standardized procedures.
- **Knowledge-based** behavior, which relates to high-level situation assessment and evaluation, consideration of alternative actions based on understood goals, followed by decision-making, planning, and execution of implementation.

For example, the application of the brake pedal on a car in response to a red light is an example of *skill-based* behavior (Wickens & Holland, 1999). However, the signals needed to trigger a skill-based response may be quite complex and multi-faceted. Wickens and Holland provide the example of a skilled emergency room physician who may immediately detect the pattern of symptoms in a patient and identify the appropriate treatment at once. A medical student presented with the same set of

symptoms may evaluate them in a much more time-consuming fashion to arrive at the same conclusion. The performance exemplified by the medical student is an example of *rule-based* behavior: an action is selected by bringing into working memory a hierarchy of rules from previous training. The decision-maker mentally scans these rules, compares them with the stimulus at hand, and selects the appropriate action. Relative to skill-based behavior, the processing is considerably less timely and automatic.
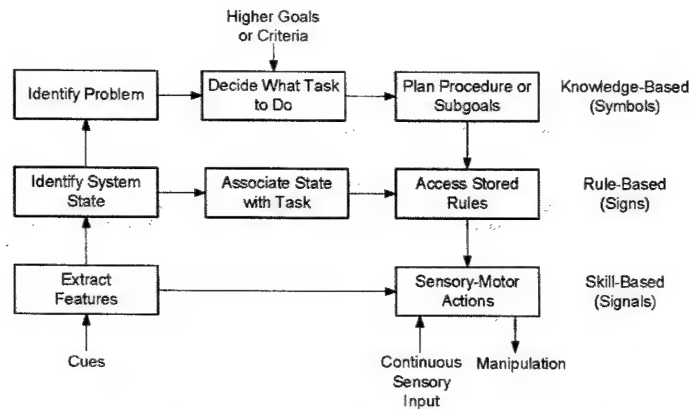
Figure 2-1: Rasmussen Hierarchy of Skilled Human Behavior

The final category is *knowledge-based* behavior, invoked when the human encounters entirely new problems for which neither rules nor automatic mappings exist. In this situation, general knowledge about the system, the goals to be achieved, and an understanding of the environment all are integrated to formulate a novel plan of action. This plan of action is then executed via sensorimotor actions.

Understanding where a given task lies in this hierarchy has implications for the effective design of an HMI intended to support that task. For example, effective executions of skill-based actions depend to a large degree on the *speed* of human response. Accuracy is important, but it is typically quite high in such tasks. Consequently, reaction time (RT) is often considered to be the critical measure of the performance quality of a person interacting with a system (Wickens & Holland, 1999). It has been found that simple RT to auditory stimuli is approximately 30 to 50 milliseconds faster than to visual stimuli (130 and 170 milliseconds, respectively) (Woodworth & Schlossberg, 1965). Speed of sensory processing between the two modalities is believed to account for the difference. Section 3.3.1 also discusses simple and choice auditory reaction times in more detail and its relationship to signal intensity. However, as we explain later in this chapter, the choice of input modality should be based on many factors other than RT. Reaction time is just one criterion; equally important may be the accuracy with which (for example) a manual tracking task can be performed.

In decision-making and diagnosis activities such as those associated with rule-based or knowledge-based behavior, *accuracy* is typically the most important performance measure. The relationship between accuracy and modality will in turn depend on the nature of the information to be conveyed, and the precision that is needed by the human.

To summarize, in choosing modalities for information presentation on a specific system, consideration should be given to whether the underlying task to be supported is by inherently a rule- or knowledge-based task (placing the premium on accuracy and precision of information presentation), or a skill-based task (placing the premium on speed and response time). However, as we discuss below, by themselves these considerations are insufficient for proper modality selection.

2.3 Compatibility of Stimulus and Response Modalities

A crucial consideration in the design of effective human/systems interfaces is the notion of *stimulus-response compatibility* in the display/control relation. This compatibility has multiple dimensions (Wickens & Holland, 1999):

- Proximity between display elements and response devices. This *location compatibility* is founded in part on the human's innate tendency to move or orient towards a source of stimulation
- Compatibility between a display and the static or dynamic properties of the human's mental model of the displayed elements
- Compatibility between stimulus and response modalities

It is the last item that is of greatest interest to us here: how do we ensure compatibility between the input modality used to present information to the human (e.g., visual, auditory, or tactile) and the response modality used by the human (e.g., manual or voice) to control a system or effect some change upon the environment?

2.3.1 Stimulus/Response Pairings

Early research by Brainard *et al* (1962) found that *choice reaction time* (i.e., the amount of time taken to make a decision as to what to do and then do it) was faster for a manual response than a voice response when the stimulus was a light. Conversely, if the stimulus was an auditorily presented digit, a vocal naming response yielded a faster choice RT than did a manual pointing response. This relates to the notion of *ideomotor compatibility* (Greenwald, 1970, 1979), which occurs if a stimulus matches the sensory feedback produced by the response. Greenwald observed fast RTs when a written response was given to a seen letter and when a spoken response was given to a heard letter. Slower response times resulted when written responses were made to a heard letter and spoken responses to seen letters. It was also found that ideomotor mappings were not influenced by the information content of the RT task nor by dual-task loading.

Teichner and Krebs (1974) conducted a meta-study to understand the factors affecting choice reaction time. Considering the four stimulus-response combinations defined by visual and auditory input and manual and vocal response, they found that:

- A manual key-press response to a light and naming of a heard digit is fastest,
- A key-press response to a digit is of intermediate latency, and

- A voice response to light is slowest

All of these research findings suggest the general notion that visual input matches well with manual output, and auditory input with speech output. However, we must take note of the following: 1) these studies largely considered reaction time as the performance metric, and thus may not generalize to rule-based or knowledge-based tasks where accuracy becomes more important than speed; and 2) many of these findings were based on single-task laboratory experiments and did not account for the effects of time-sharing and workload. We consider these issues in the remainder of this chapter.

2.3.2 The Role of Central Processing

Wickens *et al* (1983, 1984) have proposed that the concept of stimulus-response compatibility should be extended to take into consideration the type of cognitive central processing associated with the task; i.e., S-C-R compatibility. Two rationales are offered for this expansion:

- With increasingly complex systems and task environments, human operators are less likely to respond to an input signal immediately, but rather, incorporate that signal into a mental model of the system (as in the case of rule- or knowledge-based behavior). Action may be initiated, if at all, only after some delay.
- Theoretical developments from the field of cognitive psychology indicate that there are two fundamentally different *codes of representation* that underlie central processing operations of working memory. These codes may be labeled *spatial* and *verbal*.

Tasks that require spatial central processing codes are those that involving a judgment or integration concerning the six axes of translation and rotation. Examples include tracking, navigation, orientation and localization of one's own position relative to others, determining velocity vectors, or extrapolating and interpolating continuous functions. Verbal tasks are those involving the use of language, arbitrary symbolic coding, mental arithmetic, and rehearsal. Examples include communications and interaction with hierarchical data systems. Many tasks do not fit into either category, and it is more useful to think of the verbal-spatial classifications as endpoints on a continuum rather than mutually exclusive bins.

Given this classification, Figure 2-2 illustrates the potential mappings between display format/modality and working memory codes. On the input side, visual or auditory displays can be used for the presentation of either verbal or spatial information. For example, localized or 3-D audio is a means of presenting spatial information through the auditory channel, while printed text is a visual mechanism for presentation of verbal information. Experimental data collected by Wickens and others shows that the assignments of formats to memory codes should not be arbitrary: the shaded cells in the figure indicate the optimum combinations of code and modality.
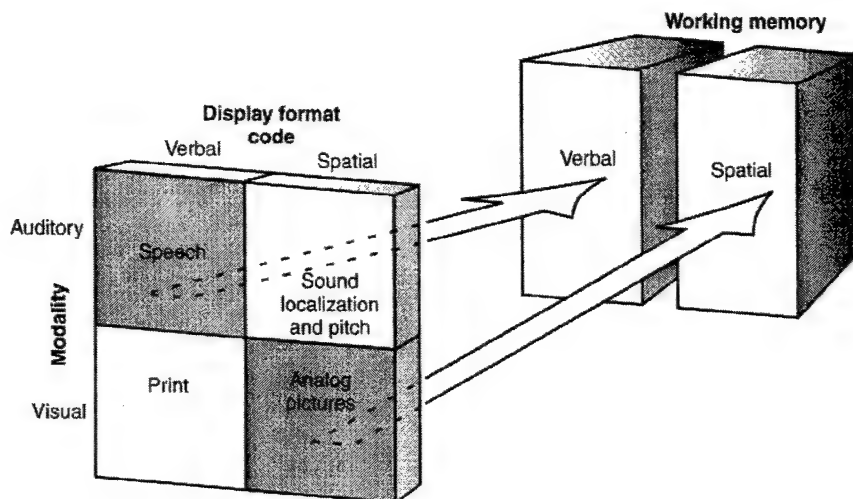
Figure 2-2: Optimum Assignment of Display Formats to Working Memory
Reproduced from Wickens & Holland (1999)

As indicated in the figure above, visual displays are most effective for tasks that demand spatial working memory, while tasks that demand verbal working memory may be better served by speech. This is especially the case if the verbal material can only be displayed for a short interval (Wickens *et al*, 1983). Their finding is supported by earlier research showing that verbal material is better retained for short periods when presented by auditory than visual modalities (Nilsson *et al*, 1977). On the response side, they propose that there exists a similar association between processing code and response modality (i.e., verbal tasks map well to a speech response, while spatial tasks map well to a manual response).

These research findings have considerable practical implications for the presentation of verbal information for temporary storage (such as navigational entries presented to a pilot). Such information is less susceptible to short-term loss when conveyed by auditory channels. However, auditory presentation becomes less effective for long messages (typically longer than four to five unrelated words or letters). In such cases it becomes necessary to prolong the message, which is accomplished more readily with print than with speech (Wickens & Holland, 1999). An optimal format may be one where auditory delivery is echoed by a permanent visual display.

It should therefore not be inferred that the off-optimal modality/code mappings are of no value; quite the opposite may be true. There exists a considerable body of research illustrating the benefits of *augmenting* primary modalities with others for the presentation of task-relevant information. For example, it has been found that adding 3D localized audio improves target acquisition performance in flight simulation tasks relative to visual-only cueing (Bolia *et al*, 1999). Along similar lines, Selcon, Taylor, and Shadrake (1992) explored the potential of multi-modal cockpit warnings. They conducted an experiment in which warning/caution visual icons and verbal warning messages were used singly and in combination to alert subjects to danger situations. The

results showed a significant decrease in response latencies when *correlated bi-modal* information was provided, as compared to either uni-modal alert. The increased information provided by two sources can increase the "recognizability" of the stimuli (through greater associational links), thus improving situation awareness and decision-making. They suggest that the presentation of correlated, bi-modal information can be a desirable design goal for functions where attentional priority is not an issue. In most real-word cases attentional priority *is* an issue, and we consider this subject in the next section.

2.4   Effects of Attention-Sharing and Multi-Tasking

Many of the studies cited earlier in this chapter considered factors such as reaction time in controlled laboratory settings where experimental subjects were presented with only a single task to perform. In most real-world task environments, a considerable amount of time-sharing may exist. In such cases, the potential for interference or competition for processing resources between concurrent tasks must be considered (Wickens *et al*, 1983).

Several theories have been developed to characterize the nature of human multi-task performance. Earlier models viewed the problem largely as one of overall resource demand and allocation. The resources needed to support multiple tasks were viewed as undifferentiated: it didn't matter whether tasks were visual, auditory, spatial, linguistic, or action-oriented (Wickens & Holland, 1999). However, it is clear that other factors to affect time-sharing efficiency. For example, it is much more difficult to read a book while driving a car than to listen to a book on tape. Using the auditory input channel for the linguistic processing dramatically changes the time-sharing efficiency of the two activities.

Research by Wickens (1980, 1984, 1991), Kantowitz and Knight (1976) and Navon and Gopher (1979) has given rise to a *multiple-resource theory* of human task performance. This theory proposes that the resources for which tasks compete may be defined by input and output modalities, by stages of information processing, and by the codes (i.e., verbal/spatial) of processing. These dimensions are thought to be somewhat independent of one another. Two tasks that share common resource demands will be time-shared less effectively than those with non-overlapping demands.

Figure 2-3 illustrates an aspect of this model of information processing in block diagram form. As shown, the resources available for perception are limited and sometimes must be shared between channels, as are the resources available for response selection and execution. Research by Pashler (1998) suggests that the latter may even be allocated on an all-or-none basis, rather than a graded one. Specifically, he found that two independent responses, based on unpredictable stimulus input, could not be selected simultaneously – one or the other was postponed. However, selection of the response for one stimulus could proceed concurrently with perceptual processing of the other stimulus.
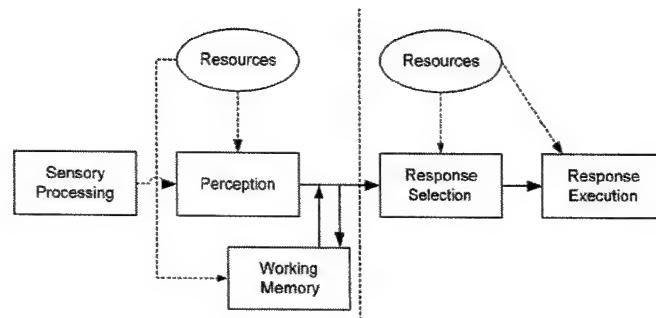
Figure 2-3: Resources Underlying Perception and Action

The distinction between the resources associated with perception, working memory, and selection of actions is illustrated in a different way below in Figure 2-4. In addition to the dichotomy between the resources for perception/working memory and response selection, differences are proposed due to auditory and visual perceptual modalities, and to spatial and verbal processing. The cube-like depiction is intended to suggest that the three dimensions are somewhat independent of one another. Operations separated by a solid line are thought to use different resources. The vertical modality dichotomy between visual and auditory resources can be defined only for perception, but the code distinction between verbal and spatial processes is relevant to all stages of processing.
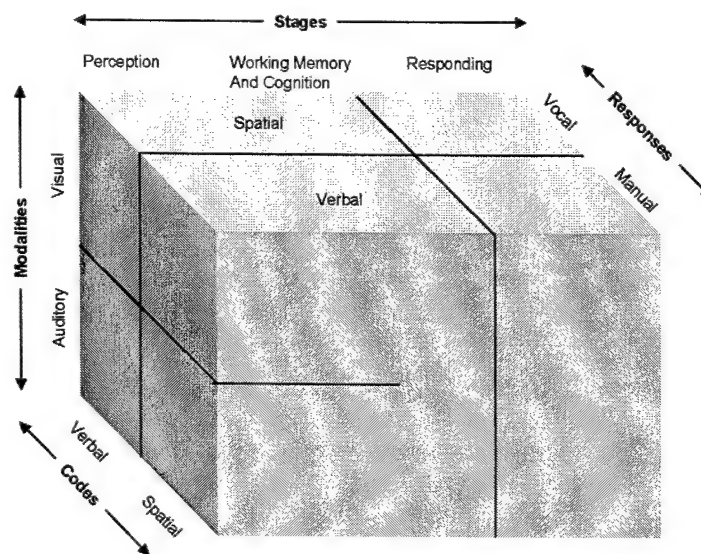
Figure 2-4: Structure of Processing Resources
Reproduced from Wickens (1984).

The resources for perceptual and cognitive processes are believed to be the same, but functionally separate from those relating to response selection and execution. This theory is based on experimental findings showing that when the difficulty of responding in a task is varied, the performance of a concurrent task making primarily perceptual demands is not affected.

With respect to the vertical (modality) axis, it has been shown that humans can sometimes divide attention between the eye and ear better than between two auditory channels or two visual channels. In other words, cross-modal time-sharing can be better than intramodal time-sharing. However, it is not clear whether the relative advantage of cross-modal time-sharing is the result of different resources in the brain being used or other peripheral factors that place audio/audio or visual/visual conditions at a disadvantage (Wickens & Holland, 1999). For example, two competing visual channels that are far apart require visual scanning between them. If placed too close together, they may cause confusion and masking, as would be the case with two auditory messages that mask one another. It has been found that when visual scanning is controlled carefully, cross-modal displays do not always yield better time-sharing (Wickens & Liu, 1988). However, in most real-word situations off-loading some information channels from the visual to the auditory modality can mitigate the dual-task interference that arises from visual scanning.

Beyond the distinction between auditory and visual modalities of processing, two different aspects of visual processing known as *focal* and *ambient vision* appear to define separate resources (Leibowitz & Post, 1982; Weinstein & Wickens, 1992) in that:

- They support efficient time-sharing
- They are characterized by qualitatively different brain structures
- They are associated with qualitatively different types of information processing

Focal vision is required for pattern recognition and resolution of fine details. Ambient vision heavily involves peripheral vision and is used for sensing orientation and ego motion. For example, reading a road sign (focal vision) while keeping one's car moving forward down the middle of the lane (ambient vision) exemplifies the parallel processing between focal and ambient vision. In the aviation community, HMI designers have considered how to exploit ambient vision for providing guidance and alerting information to pilots whose focal vision is heavily loaded.

The distinction between verbal and spatial tasks was discussed in the preceding section. Several multi-task studies have shown that spatial and verbal *codes*, whether related to perception, working memory, or response, depend on separate resources, often associated with the two cerebral hemispheres. This separation appears to explain the efficiency with which manual and vocal outputs can be time-shared (assuming that manual responses are typically spatial in nature and vocal responses are verbal).

2.5 Summary

In this chapter we have presented an overview of the relevant theories and models of human information processing that relate to the intelligent selection of human/machine interface modalities. The following topics were covered:

- A classification of **human behavior** based on the degree of cognitive processing

- **Compatibility in stimulus-central processing-response** for display/control design
- Effects of **attention-sharing** and **multi-tasking**

As discussed in section 2.2, **human behavior** can be classified into three successive levels: skill-based, rule-based, and knowledge-based. Each has a higher degree of cognitive processing and a lesser amount of automaticity than the last. Effective execution of skill-based actions depend to a large degree on the speed of human response, while accuracy is often most important for rule-based or knowledge-based activities. When choosing modalities for information presentation, consideration should therefore be given to whether the underlying task is inherently rule- or knowledge-based (placing the premium on accuracy and precision of information presentation), or skill-based (placing the premium on speed and response time).

A key consideration in modality selection is the issue of the compatibility of **stimulus-response** (SR) pairing in a display/control relation. The literature surveyed in section 2.3 indicates that visual inputs are well-paired with manual outputs and auditory inputs with speech output, when the performance criteria is a measurement of reaction time. While this provides a starting point, it does not consider the requirements of rule-based or knowledge-based tasks, nor does it account for multi-tasking effects. This notion of SR compatibility has thus been extended to consider the type of central cognitive processing associated with a task, and generalized as **stimulus-central-processing-response (SCR) compatibility**. *Spatial* tasks are those that involve a judgment or integration concerning the axes of translation and rotation, while *verbal* tasks are those involving the use of language and arbitrary symbolic coding. Spatial tasks tend to be best associated with visual inputs and manual outputs, while verbal tasks match well with auditory input and speech output. A variety of considerations constrain these findings, as discussed in section 2.3.2. In particular, the auditory modality is best used for verbal material that is relatively short in length. Furthermore, it is often the case that *augmenting* primary modalities with others can be very effective.

Finally, attention should be given to the effects of **attention-sharing** and **multi-tasking**. The multiple resource theory of human information processing proposes that the resources for which multiple tasks compete may be defined by input and output modalities, by stages of information processing, and by the codes (verbal/spatial) of processing. These dimensions are thought to be somewhat independent of one another. Consequently, two tasks that share common resource demands will be time-shared less effectively than those with non-overlapping demands.

3. Information Presentation Modalities

3.1 Introduction

In any human-machine system, efficient and effective operation relies upon constant flow of information between the machine and the human user/operator. In this chapter we describe the modalities and enabling technologies used for the presentation of information to the human. As there are five senses that allow humans to perceive the world around them, there are in theory five basic modalities by which information can be presented in a human/machine interface:

- Visual
- Auditory
- Haptic/Tactile
- Olfactory
- Taste

Most of the research and available literature focuses on the first three – visual, auditory, and tactile. These modalities are covered in sections 3.2 to 3.4, respectively. There has been some work in the area of olfactory interfaces, which is discussed in section 3.5. However, it is not an area of any significant active research, and we do not anticipate an application for this approach in the foreseeable future. The use of taste as a stimulus mechanism in human/machine interfaces is largely unexplored, and not addressed in this report. Section 3.6 summarizes the contents of this chapter.

We are ultimately interested in which modalities are best suited for conveying different types of information to humans. For example, navigating terrain requires a user to understand things such as current location, orientation of target waypoints, and distances to be traveled. Is it better to inform the user of current location by displaying the information on a heads-up display or by auditory statement via earpiece? Are bearings better conveyed through visual graphic or textual means, auditory instructions, or by vibrating the left side to indicate a left turn is needed? Understanding this research helps to create boundaries and guidelines as to how various modalities can be used successfully in systems design, how they are limited, and how they compare to one another in some instances. The way in which information is presented to the human plays a critical role in determining how well the information can be discerned, interpreted, understood, and remembered.

3.2 Visual Displays

We are inundated with visual presentations of information every day. Driving to work exposes us to a variety of visual input just within the confines of our vehicles. Speedometers inform us of how fast we are driving, iconic "idiot lights" warn us that the parking brake is still engaged, caller ID on our cell phones inform us of who is calling, the radio shows us what station we are listening to, and modern on-board navigation systems may even provide maps of our driving route. The vast area of visual displays is

far too broad to cover in this chapter. We will focus on those aspects of visual presentation particularly related to mobile or wearable applications.

Head-mounted displays (HMDs) are the leading device for mobile visual display application and form the basis for Section 3.2.1. Section 3.2.2 considers some of the issues associated with text vs. graphic presentation of visual information.

3.2.1   Head-Mounted Displays

HMDs permit an operator to view a display without looking down toward an instrument panel or other display. They come in various forms, but the common feature is that information is visually presented to the human without requiring the eyes and head to be turned toward an instrument panel or other visual display device. Other body-mounted displays have been attempted such as wrist-, arm-, and leg-mounted displays. The exact location is often driven by situational factors (how easily a location can be accessed given the tasks to be completed in the given environment), but they all encounter the similar challenge of presenting information in a visual format. Studies are somewhat mixed on the performance advantages of HMDs. While they have been used successfully in aviation, there is still concern as to their effects on user perception and attention (Glumm, 1998). The National Research Council (NRC) released reports in 1995 and 1997 reviewing the technology and the issues facing their use in "dismounted task environments." Among the NRC concerns were that HMDs might:

- Compete for soldiers' attention
- Reduce immediate situational awareness
- Conflict with performance of other critical tasks

Glumm *et al* (1998) performed field investigations to quantify the effects of HMD technology on the performance of dismounted infantry soldiers. The focus of their investigation was to compare land navigation performance using current navigational equipment to performance when using integrated HMDs. "Current" navigation equipment comprised paper map, lensatic compass, protractor, and hand-held GPS. Glumm's analysis showed that users traveled shorter distances when using the HMD system. The fact that HMDs allowed shorter routes indicates that HMDs allow more efficient navigation. One would expect times to be quicker accordingly, but times did not differ. The suggested explanation is because HMD users had to stop to consult the displays while current equipment was used en route more effectively. This seems counter-intuitive. Situational awareness was also measured according to ability to detect objects along the route and was not significantly affected by test conditions. HMDs did not significantly increase situational awareness.

Overall findings tend to indicate that the effective integration of HMD navigational information can measurably enhance navigational efficiency by providing readily accessible and easy understood position information. Although decreased travel distance did not translate into decreased time, greater efficiency should allow soldiers to be better

rested upon arrival. These results may change according to the format of display technology, information architecture, or amount of information.

3.2.2 Text vs. Graphic Information Presentation

It is important to discuss the difference between textual/numeric/written format and graphical representation. Graphical representations include icons, drawings, pictures, codes (symbols), etc. Text or numeric information can be presented in graphic form. In navigation for example, directions can be written out descriptively or graphically represented by a drawn map with icons (and perhaps textual labels or codes, combining the two formats). It is often said that a picture is worth a thousand words – if it is the right picture (Sanders and McCormick, 1993).

The general conclusion from the literature is that pictorial/graphical information is important for speed, but text is important for accuracy. The general recommendation is to combine graphics and text for speed, accuracy, and long-term retention. There is no one format that is best for numeric data. Research indicates that different formats are best for different types of information. Certain features of graphs can change the perception and lead to inaccurate interpretations. (Sanders and McCormick, 1993).

When trying to choose between symbolic or verbal format, the broad rule of thumb is to use a symbolic format only if the symbol reliably depicts what is intended. The success of a symbol relies on the strength of the association between the symbol and its referent. The association can be already recognizable or can be learned. Sanders and McCormick point out that some research suggests that symbols do not require recording, which words do. For example, a sign showing a deer conveys meaning faster than a sign with the words "deer crossing." Studies have shown reaction times to be faster for symbolic signs, especially under visually degraded condition.

This only begins to address the more specific design issues associated with creating effective graphics and codes. Those details are available in referenced text. The important point is that information can be presented visually in many formats. The usefulness or effectiveness of the information often depends upon the design of the presentation. Textual presentation may be the correct format but can easily be rendered useless if the font is too small or if there is not enough contrast against the background. Symbolic representation may be better when rapid reaction is needed, but will not work if the symbol's meaning cannot be understood.

3.3 Audio Displays

The auditory sensory modality offers unique advantages for presenting information as contrasted to the visual modality. Although it is often difficult to rely solely on audio input in most complex systems, it can offer significant benefits when used appropriately. The benefits of auditory presentation of information traditionally lead to its use primarily in warnings or alerting. Non-speech and speech-based alerting are discussed briefly in Section 3.3.1. Section 3.3.2 then presents more recent investigations of non-

traditional use of auditory presentation – particularly its use with graphic interfaces and spatial information in navigation tasks. Finally, section 3.3.3 explores why the auditory modality is a good candidate for more robust information presentation due to our ability to localize sounds. The ability to enhance auditory input via *spatialized* audio is particularly promising for its use with spatial information, which is typically a visual-only channel.

3.3.1 Audio Alerting

Sanders and McCormick (1993) suggest that the unique features of the auditory and cognitive systems make auditory presentation especially useful for signaling warnings and alarms. There is a long history of research on auditory warnings and alarms. Past literature shows that reaction time to auditory stimuli is shorter than that to visual cues – therefore, auditory signals can be useful in alerting users of critical information. More specifically, Sanders and McCormick explain that there are circumstances when auditory displays are preferred to visual displays. Auditory messages are generally preferred over visual messages when:

1) Origin of the signal is itself a sound
2) Message is simple and short
3) Message need not be accessed later
4) Message deals with events in time
5) Warnings are sent or immediate action is required
6) Continuously changing information
7) Visual system is already overburdened
8) Illumination limits use of visual displays
9) Verbal response is needed

Adams and Trucks (1976) studied reaction time to eight different warning signals in five different ambient noise conditions. Across all five conditions, the most effective signals were the "yeow" and the "beep." The least effective signal was the "wail." It is not critical to this paper to discuss the design characteristics of auditory warnings/alarms such as frequencies, tones, and duration and how that effects optimal human perception. However, it is worthwhile to note that different audio signals vary in their effectiveness for alerting functions.

It is interesting to note the relationship between reaction times and signal intensity. In simple reactions, where the same reaction is given to all stimuli, reaction times decrease with increasing signal intensity. Choice reaction time – where different reactions are needed for different signals – is more complex. Research completed by Van der Molen and Keuss (1979) indicates mid-intensity signals elicit the fastest reaction time. High-intensity signals create a startle reflex that is only useful when the same reaction is being given to all signals. The use scenario ultimately plays the most important role in determining the exact characteristics of the sound. Those guidelines are available in Sanders and McCormick but will not be discussed here.

The key understanding is that auditory modality can be an acceptable means of gaining attention and conveying warning information when the signal is purposefully designed. Some general guidelines to follow when implementing auditory warnings are:

- Auditory signals should be easily discernible from any ongoing audio input
- The same signal should designate the same information at all times
- Avoid extreme dimensions that elicit startle response
- Avoid steady-state signals by using interrupted or variable signals
- Do not overload the auditory channel at any given time

Following the earlier example of driving a car, should "idiot light" warnings be auditory instead of visual? There have been a number of efforts addressing the potential of multimodal interfaces in cockpits. Selcon, Taylor, and Shadrake (1992) conducted an experiment in which visual warning icons and verbal warning messages were used singly and in combination. Results showed significantly faster response times when bimodal warnings were given as compared to either unimodal alert. This suggests that correlated bimodal presentation can be desirable for functions where attention priority is not an issue (Mulgund and Zacharias, 1996).

Speech displays should be considered in addition to simple auditory warnings and coded messages. Advantages of speech over other auditory signals include:

- Flexibility
- Ability to identify the message source
- Not coded (codes require training and can be forgotten under stress)
- Rapid, two-way exchange of information

One ongoing debate in the design of speech displays is the message format and the use of auditory alerts preceding the speech message. The major issues relate to the effectiveness of monosyllabic versus polysyllabic words, keywords versus sentences, and speech messages with or without alert tones. In both natural and synthetic speech, polysyllabic words are more intelligible than words monosyllabic words. Similarly, words in sentences are more intelligible than words in isolation. The context provided by additional syllables and words help to reduce the ambiguity of the meaning (Blackwood et al, 1997). There is conflicting research as to how and when speech warnings/messages should be used. Some advocate speech for less time critical tasks and implement sentences when disruptions are possible or when there are a large number of alternatives. Others think speech is only suitable for time-critical warnings.

Table 3-1 provides guidelines for the use of different auditory displays. The three types of displays are simple tones, complex non-speech sounds, and speech.

Table 3-1: Auditory Display Format by Function
Reproduced from Blackwood (1997)

Function	Tones (periodic)	Complex Sounds	Speech
Quantitative indication	POOR Maximum of 5 to 6 tones absolutely recognizable	POOR Interpolation between signals inaccurate	GOOD Minimum time and error in obtaining exact value in terms compatible with response
Qualitative indication	POOR/FAIR Difficult to judge approximate value and direction of deviation from null unless presented in close temporal sequence	POOR Difficult to judge approximate deviation	GOOD Information concerning displacement, direction, and rate; presented in form compatible with response
Status indication	GOOD Start and stop timing; continuous information where rate of change is low	GOOD Especially suitable for irregularly occurring signals (alarms)	POOR Inefficient; more easily masked; problem of repeatability
Tracking	FAIR Null position easily monitored; problem of signal-response compatibility	POOR Required qualitative indications difficult to provide	GOOD Meaning intrinsic to signal
General Comment	Good for automatic communication of limited information Meaning must be learned Easily generated	Some sounds available with common meaning Easily generated	Most effective for rapid (but not automatic) communication of complex, multidimensional information Meaning intrinsic to signal and context. Minimum training and learning required

3.3.2 Auditory Presentation of Visual Information

Graphical user interfaces and spatial information are traditionally presented using visual means. Visual interfaces have benefited many applications, yet they present exclusive barriers for those that are visually impaired and those who cannot rely on sight in such cases as darkness, etc. Creating access to visual information in nonvisual ways is a growing concern as designers try to achieve "universal design." Auditory presentation is one nonvisual modality that offers some promise for presentation of graphical and spatial information.

3.3.2.1 Graphics

Creating auditory access to interactive graphical interfaces such as a desktop PC presents unique challenges. Simple approaches like screen-readers can be used to translate graphics into more basic interactions, but it is advantageous to have the nonvisual access parallel the visual access so that non-sighted and sighted users can work together on the same program at the same time. Two projects, Mercator and

GUIB, explored this issue using two different approaches. Georgia Institute for Technology developed Mercator; a consortium of European partners developed GUIB (Textual and Graphical User Interfaces for Blind People).

The GUIB design is primarily based on tactile interaction but also contains some nonspeech auditory cues. Using an input-output device called GUIDE that integrates Braille, loudspeakers, and a touch tablet, GUIB translates screen contents into tactile presentation based on the spatial organization of the graphical interface. Mercator, on the other hand, utilizes a hierarchical auditory interface with speech and nonspeech auditory cues used to convey iconic information. Mercator uses "auditory icons," which are sounds associated with daily objects that correspond to the graphical icon. Touching a window, for example, sounds like tapping on glass and editable text fields sound like an old typewriter. This approach is easy for some interface objects like trashcans and windows but is more complex for abstract items such as menus and dialog boxes. User studies showed that auditory icons are easily learned but initial use suffers from frustration with identifying cues. Both GUIB and Mercator projects provide valuable experience in understanding how to translate graphical interfaces into nonvisual presentation. Interestingly, both have plans to move toward multimodal presentation, using both nonspeech audio cues and tactile Braille cues (Mynatt and Weber, 1994).

Supporting this effort toward integrated audio and tactile presentation of visual information, a new computer mouse was developed that has both tactile and audio representations of graphs. The mouse vibrates every time it meets a line on a graph, and tones vary in pitch according to whether the line is rising or falling (CNN Reuters 9/9/02).

3.3.2.2 Spatial Information

As discussed in Section 2, spatial information relates to tracking, navigation, orientation, localizing one's own position relative to others, determining velocity vectors, or extrapolating and interpolating continuous functions. The presentation of spatial information traditionally calls for visual presentation. Nonvisual presentation of spatial information, however, is of high concern for the visually impaired community as well as sighted users whose visual workload may already be too high. While auditory modalities may sometimes serve as primary information sources, it may also be useful to use the auditory channel to offload information from the potentially over-burdened visual system. Information overload can be as detrimental to performance as a poorly designed or unreadable display (Glumm, 1999). This section will review research involving auditory presentation of spatial information in navigation tasks.

Glumm (1999) studied the effects of an auditory presentation versus a visual presentation on soldier navigation performance and target acquisition tasks. In auditory mode, position information was presented via verbal messages; in visual mode, the same information was presented in text and graphic form on a map using an HMD. Subjects accessed both modes through a belt-mounted keypad. The frequency at which navigational and other tactical information was accessed was the same for both visual

and auditory presentations. There was no difference found in terms of navigation and target acquisition performance. The findings did indicate that soldiers maintained greater awareness of position (waypoints, targets, and other units) when information was presented visually on an HMD than when the same information is presented audibly by verbal messages. Some reasons for this include:

- Auditory position information must be presented in series rather than in parallel, and translating that into mental images is thought to be more difficult.
- These results may be attributed to factors affecting ability to retain and recall information. Research and literature about memory indicate that information must be rehearsed in order for it to be retained. Even with rehearsal the information can decay over time, and decay is more rapid with more information trying to be kept in short-term memory. Other research shows that a visual image stored in working memory will decay faster than an auditory image, but the capacity of the visual image store is larger than that for auditory image and visual images can often be referred to more easily.

Coming back to the driving example, we find another instance where users may require navigation information while their visual channel is occupied with tasks related to primary driving task. Back Seat Driver is a navigation system developed at MIT Media Lab for taxi drivers. Rather than having to look at maps or manually input queries, the system provides directions through a speech synthesizer and responds to voice commands. No results were available in terms of its performance.

It is important to note that in many "real world" scenarios, humans will be dealing with greater amounts of information than is seen in refined scientific studies. That information is all competing for the same limited cognitive resources. As the amount and diversity of information increases, so will menu and visual display complexity. The design of visual and auditory displays and proper use of these modalities for information presentation will have a major impact on performance (Glumm, 1999). There is a need to better understand the interaction between auditory and visual modes while much of existing literature treats them independently.

Again, consider our driving example. Speech recognition and text-to-speech enable new driver-vehicle interfaces that could potentially increase safety, allowing drivers to keep two hands on the wheel. It is sometimes assumed that this auditory-based interaction will not distract the driver, but this ignores the cognitive load of speaking and listening and its effect on driving performance. Comprehensive reviews suggest that speech-based interactions have the potential to distract drivers and degrade safety. "Conversing" with a computer has the added demands of having to interpret synthesized voice and having to navigate menu structures of on-board computers. Lee *et al* (2001) investigated the effects of speech-based email on drivers' attention. They found a 30% increase (310ms) in reaction time (to braking for leading vehicles) when the speech-based system was used. Subjective response indicate that speech interaction introduced a significant cognitive load, increasing with complexity of the email interface.

Research was done to examine auditory access to spatial information for visually impaired travelers in unfamiliar environments. Loomis *et al* (1994) compared two auditory displays as part of a navigation guidance system. A conventional verbal/speech display through earphones was compared to a "virtual acoustic display" using binaural earphones. The virtual acoustic display indicated landmarks by having their labels/locations presented as virtual sounds at the correct locations within the space. Initial testing of this navigation system was simple and only required walking in linear segments featuring two endpoints and numbered waypoints along the segment. There was a period of acclimation needed, after which no difficulty was experienced while orienting toward the target endpoint and subsequently walking toward that landmark.

Auditory interactions are even opening up opportunity for more universal access to video gaming. ZForm is attempting to make computer games with advanced audio so good that blind people can play. They are developing a complex version of the computer game Quake. Quake is a "first-person shooter" game that requires players to navigate through maze-like buildings and environments, choose pathways, open doors, locate and pick up equipment from the floor – all while fighting enemy forces. The team created a world of aural cues to replace the intense visual scenes presented to sighted users. When a hallway opens on the left, a slight rushing of air is heard to the right. Weapons give off characteristic sounds that are spatialized to indicate their location (this concept is further discussed in Section 3.3.3). One of the ZForm cofounders, who is blind, recently competed against sighted players using the auditory interface and was not only able to compete but was also able to defeat sighted players using the standard visual interface (Cook, 2002).

3.3.2.3 Speech Pattern Effects

Often, auditory input/presentation is coupled with speech output. Users hear information from the system and speak information back into it. Because there is an added cognitive burden created by auditory access to spatial data, however, speech patterns (back to the system) were expected to change in those instances. Baca (1998) investigated the impact of auditory access to spatial data on the nonverbal aspects of speech – such as pauses and intonation. Experiments with navigation tasks attempted to study this by comparing auditory conditions with multimodal conditions. After listening to verbal descriptions of overall layout of the environment, subjects were told a starting point and a destination. They then used the application to plan a route to the given destination. Subjects were required to do so using an auditory interface and a multimodal interface which featured a touch screen and map along with the speech query.

Statistical analysis showed a significant difference in nonverbal aspects of spoken language when using auditory versus multimodal interface. Participants universally preferred to use a mixture of fixed natural language commands and freely formed natural language queries. They preferred fixed commands for events such as navigating along a route ("go forward," "go back"); preferred freely formed queries when asking for information from the database ("what is the traffic like on this road?"). There was strong evidence in the data that hesitation pauses at locations other than natural phrase

boundaries are increased in the auditory condition for all categories of users. These pauses indicate an increase in cognitive effort when trying to use the auditory modality for navigation tasks. Additionally, these pauses tend to increase the error rate of the recognizer, which impacts system performance and user satisfaction.

3.3.3 Spatialized Audio

The ability to identify the direction from which sound waves are emanating is called stereophony and is driven primarily by the difference in intensity and phase of sound reaching each ear. When sounds are coming from peripheral locations, each ear receives the sound at slightly different times. The brain interprets the phase difference to determine direction/location. This is referred to as sound spatialization or localization.

Caelli and Porter (1980) investigated a practical example of sound localization – determining the direction of an ambulance siren. Subjects sat in a car and heard a 2-second burst of a siren 100 meters from the car. Subjects tended to overestimate the distance (often by a factor of 2), thinking the ambulance was farther away than it was. With respect to direction, estimates were often 180 degrees off, particularly when coming from behind. Performance was also negatively impacted by open driver's side window as added noise on one side affects how each ear heard the sound of the siren.

Carrying sound localization a step further, Strybel (1988) explored the ability of people to detect the motion of a sound. The measurement used is the minimum audible movement angle (MAMA), defined as the minimum angle traversed by a sound source which enables a listener to detect motion. The MAMA depends upon the speed of the motion and the initial position of the sound. Using simulated sound, Grantham (1986) found a 5 degree MAMA with initial position straight ahead; MAMA of 33 degrees when initial position was moved to side. Using real noise source, Manligas (1988) found a MAMA of 1 degree for straight ahead and 3 to 4 degrees when initial position was 80 degrees left or right. Strybel provided a few conclusions:

- The auditory system is relatively insensitive to motion displacement when compared to the visual system
- Detection of simulated sound movement is worse than detection of real sound
- Detection of movement of broadband noise is easier than pure tone

Abouchacra (2001) studied the effects of spatialized sound presentation on the listener's ability to detect "target messages" among competing messages at high-level background noise, simulating a military environment typically found in vehicles for example. The use of spatialized audio improves speech recognition. The best overall performance (in terms of message detection) was found with spatialized presentation. In general, speech recognition was shown to be better in quiet modes than in background noise conditions with competing messages – as one would expect. The work of Abouchacra supports the claim that spatializing messages in acoustic space can improve auditory performance when there is competition for resources, even in high noise levels.

Spatialized audio becomes most helpful, perhaps, when trying to create auditory access to spatial information (as discussed in Section 3.3.2). Bolia *et al* (1999) conducted an experiment to evaluate the effectiveness of spatial audio displays on target acquisition performance. Subjects performed visual search tasks with and without the help of spatial audio cues, under varying distracters. Results indicated that both free-field and virtual audio cues result in decreased search times (faster target acquisition).

3.4 Haptic/Tactile Interfaces

The field of haptics addresses information acquisition and object manipulation through the sense of touch. *Haptic* means "of or relating to the sense of touch." In this section we consider the use of touch for conveying information *to* the human; the subject of manipulation and control is addressed in the next chapter. Biggs and Srinivasan (2002) summarize the categories of applications for which haptic interfaces offer potential to support human task performance:

- **Medicine**: surgical simulators for training, remote diagnosis for telemedicine, and aids for the disabled
- **Entertainment**: video games and simulators that enable the user to feel and manipulate virtual solids, fluids, and avatars, and receive physical feedback in response to events happening in the simulated world
- **Education**: enabling students to get the "feel" of physical phenomena at nano, macro, or astronomical scales; experiencing complex data sets
- **Industry**: integration of haptic feedback into CAD systems, allowing the designer to manipulate the components of an assembly in an immersive virtual reality environment

Considered against the cognitive task classification presented in the previous chapter, a common element of most of these tasks is that they are spatial in nature. While haptic interfaces can be used for simple alerting functions (e.g., vibrating pagers) that cue the human to some condition of interest, the applications listed above entail a more involved physical interaction between the human and some system. The haptic interface offers the potential to improve the user's sense of presence in the task environment.

3.4.1 The Human Haptic System

When a person touches an object, forces are imposed on the skin. The net forces experienced, together with the posture and motion of various limb segments are conveyed to the brain as *kinesthetic* information. This information is conveyed by multiple sources such as receptors in the joints, muscles, and tendons (Biggs & Srinivasan, 2002). This is the means by which humans sense the coarse properties of objects. The spatial and temporal variations of force distributions within the contact region on the skin are conveyed as *tactile* information by several types of receptors embedded in the skin. Attributes such as fine texture, softness, and slipping of surfaces are perceived through the tactile sensors. Skin temperature (which is related to the

temperature of an object being touched) is also sensed through specialized tactile sensors.

The human haptic system also consists of the motor system enabling active exploration or manipulation of the environment, and a cognitive system that links sensations to perception and action (Biggs & Srinivasan, 2002). A *tactual image* consists of both tactile and kinesthetic sensory information, and is controlled by motor commands reflecting the user's intention. Broadly speaking, haptic interfaces can be viewed as generators of mechanical impedances that represent a relationship between forces and displacement (and perhaps their derivatives) over different locations and orientations on the skin surface. In many simple tasks involving active touch, either tactile or kinesthetic information is fundamental for identification and discrimination, and the other is supplementary. For example, kinesthetic information is fundamental for the determination of the length of rigid objects held between the thumb and forefinger (Durlach *et al*, 1989). These tasks require sensing and control of variables such as fingertip displacement. In contrast, tactile information is fundamental for detection of surface texture or slip. In this case, sensing the spatiotemporal force distribution within the contact region provides the basis for making inferences about object properties. Both types of information become necessary and equally important in more complicated haptic tasks. There exist numerous specific examples of the use of both tactile and kinesthetic cueing, and we discuss some of them in the following two subsections.

3.4.2 Tactile Displays

At the Haptic Interface Research Laboratory at Purdue University, a tactile directional display has been developed by embedding a 3x3 array of tactors into both the back of a chair and a vest (Tan, Lim, & Traylor, 2000). This technology takes advantage of the *sensory saltation phenomenon*, which is a haptic spatiotemporal illusion that can evoke a powerful perception of directional lines. The sensory saltation phenomenon was discovered in the early 1970s. In an experiment leading to its discovery, three mechanical tactors were placed equidistantly on the forearm of the test subjects. Three brief pulses were delivered to the tactor closest to the wrist, followed by three more at the middle position, and then three more pulses at the tactor farthest from the wrist. Rather than sensing three successive taps localized at the tactor sites, subjects were under the impression that the pulses were distributed uniformly from the site of the first to that of the last tactor.

Tan *et al* summarize the constraints on tactor placement and configuration under which this illusion is effective. The objective of their research was to design a back display based on sensory saltation for providing tactile directional signals. It was envisioned that such an interface could be useful in scenarios where visual or auditory information are unavailable or obscure, and directional signals are needed to support a particular task. A vibrotactile array with inter-tactor spacing of 8 cm was sewn between two supporting layers of fabric, so that it could be draped over the back of an office chair. Their experimental results with untrained subjects demonstrated a consistent interpretation of

the test signals as indicators of direction, and suggest the potential of using this approach for a general-purpose haptic directional display.

Rupert (1998) suggests that tactile displays offer considerable potential to combat spatial disorientation in aviation environments. He notes that the U.S. Army has observed an increase in the number of spatial disorientation mishaps since 1985, coinciding with the increased use of night vision goggles. This technology made possible mission profiles such as nap-of-the-earth flight, night formation flight, and all-weather flying. All of these conditions introduced new opportunities for the pilot to experience spatial disorientation.

The Tactile Situation Awareness System (TSAS) was developed to combat this disorientation problem, by providing an intuitive tactile display for continuous communication of true orientation information. The TSAS is a matrix of tactile stimulators incorporated into a flight suit, designed to provide the pilot with critical flight parameters via the sense of touch. It is designed to communicate with the pilot at the level of "reaction" and "subconscious" behavior, so that interpreting its output does not require much conscious effort. The TSAS provides tactile stimuli to the areas of the torso where a pilot would normally receive pressure cues on the ground, were the pilot firmly attached to a chair with multiple straps. Rupert reports that training a user (whether an experienced or novice pilot) takes only minutes, because the system's logical operation is consistent with their mental model of spatial orientation based on cumulative life experience. It is believed that presenting spatial orientation information in a manner that does not place a cognitive load on the pilot will increase the resources available to attend to other tasks.

3.4.3   Kinesthetic Cueing

Force feedback systems are becoming increasingly common in applications such as flight simulation, where they are used to provide simulated force cues to the pilot. While originally in the domain of high-performance, high-cost training simulators, they are now available commercially for use with off-the-shelf PC video games.

Research sponsored under DARPA's *Augmented Cognition* program has studied the use of kinesthetic cues as an aid to human memory. Tan *et al* (2002) found that kinesthetic cues (i.e., the awareness of the parts of the body's position with respect to itself or to the environment) were useful for recalling the positions of objects in space. Their experimental study found a 19% increase in spatial memory for information controlled with a touchscreen (which provides direct kinesthetic cues), as compared to a standard mouse interface. The objective of their research was to develop human/machine interface concepts that make information more memorable and easily recalled. The results suggest the potential of using kinesthetic cueing as a mechanism for enabling effective recall.

These findings are consistent with research demonstrating the benefits of *quasi-modes* in human/machine interfaces (Raskin, 2000). Modes have long been a significant source of

errors, confusion, and complexity in human/machine interfaces. Raskin constructs a definition of interface modes in terms of *gestures*: a gesture is a sequence of human actions completed automatically once set in motion (e.g., pressing some combination of buttons or keys, sliding a control, etc.). Modes manifest themselves in terms of how an interface responds to gestures. For any given gesture, the interface is in a particular mode if the interpretation of that gesture is invariant. If the gesture has a different interpretation, the interface is in a different mode. Mode errors can arise when the state of the interface (and therefore its current mode) is not within the user's locus of attention.

What, then, are quasi-modes? Consider the use of the Caps Lock key to type uppercase letters. This is very different from using the Shift key for the same effect. The Caps Lock key establishes a mode in the interface, while the Shift key does not. It has been shown (Sellen *et al*, 1992) that the act of holding down a key, pressing a foot pedal, and any other form of physically maintaining an interface in a particular state does not induce mode errors. There exist neurophysiological roots for this phenomenon. Much of the human nervous system operates in a way such that a constant stimulus yields signals that decrease over time in their ability to capture attention. This continues until the cognitive system receives no signal at all. However, signals that report whether muscles are actively producing a force do not fade. The term quasi-modal can thus be used to denote system modes that are maintained kinesthetically.

In this context the distinction between providing feedback to the user and providing a mechanism for manual control is blurred; however the key observation is that using kinesthetic cueing as a component of a user interface can enhance usability and reduce the potential for mode errors.

3.5   Olfactory Interfaces

Olfactory interfaces remain one of the least developed areas in human/computer interaction (Youngblut *et al*, 1996). Many people think that the sense of smell, because of its sensitivity, is a good sensory modality. This sensitivity depends on the particular odor and the individual, and while we are good at detecting the presence of an odor, there is also a high false-alarm rate. Our sense of smell is not good at making absolute identification; we do far better when comparing odors in a relative manner. Research regarding odor identification, training, vocabulary, and odor intensity indicates that smell is better used to determine the presence of an odor but should not be counted on to identify a specific stimulus (Sanders and McCormick, 1993).

While considerable progress has been made in the development of electronic noses that can acquire and interpret odor data, systems that can provide olfactory cues to a human have met with less success. There is considerable evidence that odors can be used to manipulate mood, decrease stress, increase vigilance, and improve retention and recall of learned material. As with auditory cues, the potential exists to use odors for *sensory substitution*, to represent phenomena that have no smell or are purely abstract information.

Several components are needed for the implementation of an olfactory delivery system. These include:

- Odor storage and display
- Odorant selection
- Cleaning of air input
- Evacuation and cleaning of exhaled air

Storage of odorant may be the most mature of these. Odorants can be stored in several ways, including as liquids, gels, or waxy solids. A common method in research systems tends to be microencapsulate odorants, which is the basis for scratch-and-sniff patches. Major delivery methods include air dilution olfactometry, breathable membranes coated with a liquid odor, and liquid injection into an electrostatic field with airflow control. Table 3-2 summarizes the various delivery technologies that have been used.

Table 3-2: Technologies for Olfactory Delivery
Reproduced from Youngblut *et al*, 1996.

Storage Technologies	Presentation Technologies	Advantages	Disadvantages
Liquid	Unpowered evaporation Saturated cotton balls Breathable membranes Permeation tubes Bubble chambers	No power required Inexpensive	Bulky Odorants are clumsy to handle
	Heat-induced evaporation	Inexpensive	Power-hungry
Gels	Electrostatic evaporation	Good for large spaces Materials easier to handle	Never miniaturized Requires higher voltages
Microencapsulation	Mechanical Release	Could be valveless Materials easy to handle	Mass production technology Impractical for small lots
	Heat Release	Could be valveless Materials easy to handle	Mass production technology Impractical for small lots
	Valve Design Options:		
	No valves	Smaller, cheaper	Inter-contamination of odors
	Off-the-shelf valves	Mass-produced	Bulky, power hungry Fast or precise, not both
	Ink jet printer nozzles	Precise control	Single units large because of packaging
	Microvalves	Potentially fast & small	Must make custom manifolds to obtain greatest miniaturization

In addition to storage and delivery, olfactory interfaces must also clean the air input, select which odorants to display, and evacuate and clean exhaled air. Controlling breathing space for the individual is the great obstacle: odor intensities must be controlled accurately and flushed from the breathing space when no longer needed, and it must be ensured that no contamination occurs from persistent odors. Krueger (1995) summarizes several ways for presenting odors:

1) A sealed room with a precise air filtration system
2) An unsealed cubicle that directs treated air toward the user's face, and evacuates odorized, exhaled air with a vent behind the user's head
3) A completely sealed pod in which the user breathes only treated air, and exhaled air is evacuated continuously
4) A tethered mask that is usable in any room by either a seated or stationary standing user
5) An untethered system consisting of a belt pack and tubes running to and from a mask in an HMD
6) An untethered system that is completely incorporated into an HMD

Most of these concepts were contemplated for use as olfactory cueing systems in virtual environments. Only (5) and (6) would be applicable for use by dismounted infantry. In addition to the amount of encumbrance to the user, each of these techniques varies considerably in terms of cost, space, and support requirements.

Youngblut *et al* (1996) summarized a number of commercial efforts (at that time) to develop systems for olfactory cue delivery. These included products by the BOC Group plc, Ferris Productions, Inc., Marketing Aromatics Ltd., and the Artificial Reality Corporation (ARC). ARC's work to develop olfactory interface technologies for virtual environments was funded by DARPA, but discontinued three years ago for lack of interest in the marketplace (Krueger, 2002). *Wired* Magazine recently described the efforts by a startup company called Digiscents to develop an olfactory cueing system for personal computers (Manjoo, 2001). Their *iSmell* product was touted as a "personal smell synthesizer." That company no longer appears to exist. In fact, no recent information could be found about any of the aforementioned systems. There still does not appear to exist any commercially viable product for controlled olfactory cue delivery.

Given the technical challenges associated with practical olfactory cueing systems and the observation that none of the efforts highlighted so prominently just a few years ago have met with great success, it seems unlikely that this modality will be a useful component of real-world HMIs in the foreseeable future. The most promising use of smell remains as a warning device such as gas companies adding odorant to natural gas so that we can detect gas leaks and mines relying "stench" systems to warn miners in an emergency. While olfactory displays are unlikely ever to become widespread, they offer an interesting and unique form of display that might one day be used to supplement more traditional interfaces (Sanders and McCormick, 1993).

3.6 Summary

We have presented an overview of the different modalities that can be used for information presentation in a human/machine interface. The modalities follow the five senses that humans use to detect signals – visual, auditory, haptic, olfactory, and taste. Practical applications are limited to the first three. While in principle all of these options exist for use in an HMI, system designers should not use them haphazardly: each one

has strengths and weaknesses in dealing with different types of information. As discussed earlier in Chapter 2, another important consideration is the type of task that the system must support. There are no steadfast rules, but the three tables below outline the applicability of the information presentation modalities discussed in this chapter.

Table 3-3: Applicability of Visual Displays

Technology or Methodology	Applicability
Head-Mounted Display	• Overall, visual presentation is applicable for spatial information • Visual textual displays acceptable for verbal information, particularly longer messages • HMD particularly useful in situations where information is needed without diverting gaze, "on-the-move" operations • Research suggests HMD increases efficiency in navigation tasks which are mobile and spatial • Inconclusive evidence regarding attention and situational awareness
Text vs. Graphic Presentation	• Graphic/symbolic information preferred for speed; effective only if symbol depicts intended meaning via associations • Text information preferred for accuracy • Graphics and text can be used together for speed, accuracy, and long-term retention • Visual display of text can be coupled with gesture-based sign language applications to show the signed message, eliminating some time and cognitive load needed for interpretation

Table 3-4: Applicability of Auditory Displays

Methodology	Applicability
Alerts/Warnings	• Auditory displays particularly useful for signaling warnings, alarms, or critical information because of fast reaction times (exact design of signal intensity, tone, and variation impacts effectiveness) • Auditory alerts or messages should be short, simple, and not needed repeatedly • Auditory signals good for attention when immediate action is needed or when information is continuously changing • Synthetic speech warnings instead of tones offer flexibility, no codes to learn, and fast two-way communication • Auditory messages may be paired with verbal response (speech control) • Bimodal (visual with auditory) warnings can elicit faster response than either used alone
Spatial Information via Audio	• "Auditory icons" can be effective for presenting graphic information in auditory format • Auditory access to graphic information is moving toward multimodal w/ tactile manual devices • Audio presentation of spatial information in navigation tasks (retain/recall diminishes compared to visual presentation – auditory may only be useful for instructions like "turn left here") • Cognitive demands of auditory-spatial can increase reaction times and complexity of operation
Spatialized Audio Cues	• Audio signals can be spatialized to indicate direction and location and movement, increasing its effectiveness for spatial presentations • Spatialized 3D audio can help identify auditory messages in noisy conditions or could be used for navigation tasks (waypoints, object/person locations, etc) • Detection of simulated sound is worse than that of real sound, but could improve with technology

Table 3-5: Applicability of Haptic Displays

Methodology	Applicability
Tactile	• Tactile presentations are effective for simple alerting via vibrations, pressure, etc. • More complex applications to provide sense of presence, orientation, or direction in a task environment without added cognitive burden
Kinesthetic	• Kinesthetic cues not usually used alone • Kinesthetic cues can supplement other displays to help remember location of items in space relative to self, increasing recall • Increase usability and reduce mode error by designing modes that are maintained kinesthetically

The effectiveness of all modalities ultimately relies on display design and concurrent physical and cognitive demands. These are guidelines to begin understanding how each modality may best be used for presenting information. As technologies continue to advance, system designers can draw upon multiple modalities for intuitive interactions that go beyond the traditional approaches to human/machine interface design, which have long centered on visual displays combined with manual controls.

4. Control Modalities

4.1 Introduction

In this chapter we discuss the modalities and enabling technologies that allow humans to control their systems in some way. In recent years there has been an explosion in the number and capability of devices to facilitate interaction between humans and machines. Ideally, to make this interaction as natural as possible, computers should be able to interpret all natural human actions – hand and body movements, facial gestures, speech, eye gaze, etc. Some more recent studies are going a step further to see how machines can be controlled simply by thinking about actions. Section 4.2 reviews manual control, which is the more traditional and most exploited modality. Gesture-based controls are discussed as a subset of manual control. Section 4.3 considers both head and eye gaze-based controls. Section 4.4 reviews speech interfaces and various issues surrounding the use of speech-based interactions. Section 4.5 reviews recent explorations in neural control. Each of these sections deals with the modality primarily by itself. Section 4.6 then explains multimodal control systems and the rationale for using them. Finally, section 4.7 summarizes the contents of this chapter.

4.2 Manual Control

There are many interactive manual modes that can be used to link human response to a machine/system action. Large, slow controls requiring large amounts of force can use the arms and legs for strength such as with levers or foot pedals. Smaller controls call upon hand and finger actions for quick, accurate, coordinated control with switches, toggles, dials, buttons, and keypads (Adams, 2001). Historically, these movements of the hand are the actions most exploited for HCI. The human hand offers dexterity and the ability to apply appropriate force and acceleration – all good for positioning and controlling devices. A number of simple manual devices have been designed to capture these benefits, including the keyboard, mouse, stylus, pen, wand, joystick, trackball, etc. (Sharma, 1998). These are discussed in Section 4.2.1.

Another level of control involving the hands (and sometimes other body segments) is *gesture-based control*, where the body itself becomes the control device. Gesture-based controls are used in many of the same manipulation scenarios but offer greater flexibility as one can grab and turn a virtual object as if they are holding the object in reality. Gesture-based systems are also being used for sign language applications. Section 4.2.2 discusses gesture-based control. We have included gesture-based controls as a subsection of manual control because literature often overlaps these areas.

4.2.1 Simple Manual Devices

Overall, there are several classes of manual controls. Adams (2001) outlines ten general categories of manual controls including:

1) Large linear controls (pedals, levers, cranks)
2) Small linear controls (bars, slides, push buttons)
3) Switches (toggle, rocker)

4) Large rotary controls (cranks, wheels, yokes)
5) Small rotary controls (cranks, thumb wheels, knobs)
6) Keyboards
7) Mouse
8) Touch devices (touch screen, panels, membranes)
9) Joysticks
10) Tracker ball

Because we are concerned with the requirements of more mobile systems, we will limit the discussion of this section to the smaller controls that may be featured in a wearable system. These types of manual controls are used for a variety of applications, including on/off functions, text editing (word processing), spreadsheets, drawing programs, more advanced computer-aided design programs, and other precise system controls. There are numerous human factors concerns associated with the devices in the following sections (such as dimensions and actuation forces), but we will not cover those in detail here. That information can be found in various human factors standards and text books. The following subsections will briefly describe various manual devices, how they are typically or best implemented, and some recent technical advances within each.

Note that the literature on manual controls is heavily weighted toward the human factors aspects of control selection and design. There was not an abundance of information found about the performance of these devices, or comparisons between them in terms of handling tasks.

4.2.1.1 Keyboards

Keyboards are most popular and valuable for capturing strings of text. While speech recognition is becoming more and more available, many think that the keyboard will remain widely used for this purpose. Typing is the most common way of entering information into a computer because it is reasonably fast, very accurate, and requires no computational resources (assuming standard two-handed keyboards are used). Some of the downsides to keyboards are that they require training and practice, they lack standardization across different keyboard layouts, and they have a limited ability to manipulate graphical interfaces. One-handed keyboards are being developed and raise some concern because of learning times involved and the added cognitive load of having to remember various finger combinations.

A new technology called Multitouch is being explored as a replacement to the conventional electromechanical keyboards. Multitouch is a thin sensor array that recognizes your fingers and hands as they move over the surface. They are not pressure sensors – minimal contact is required, as keys do not need to be depressed. Users get auditory feedback to indicate successful contact, if desired. Cursor positioning can be done (instead of using a mouse) using the same surface, but with two fingers instead of one used for typing. The same surface can be used to capture handwriting and graphical input as well (Hedge, 2002).

4.2.1.2 Mouse and Trackball

The mouse was developed at Stanford Research Laboratory (now SRI) in 1965 to be a cheap replacement for light pens (Myers, 1996). The mouse offers an improvement over keyboard for dealing with graphic interfaces, particularly for moving and selecting objects. Its drawback is that it presents an additional center of focus beyond the monitor and, often times, an accompanying keyboard or other device. Adams (2001) also explains that the mouse is not suitable for drawing because it is moved with the wrist and arms, which lack the finer controls of the hand and fingers.

There is no conclusive evidence favoring the mouse or trackball for cursor movements. In general, the preference is based on space limitations. Mice are used where there is room for the contact pad; trackballs are used for confined operations (Adams, 2001).

4.2.1.3 Touch Screen and Pen

Touch screens and pens can be used together or singly. Modern touch screens can range in capability from simply pointing and selecting objects such as an ATM interface to being able to recognize writing such as many PDA's. Writing capture is sometimes dealt with in the literature as gesture-based control because the computer is interpreting the gesture of writing, but we consider this to be a simpler manual action than what is discussed in section 4.2.2. Pens have been used for decades for a variety of tasks. Early graphics interfaces relied in light pens for manipulation before the days of the mouse (Myers, 1996).

Cambridge Consultants Limited recently developed a pen that enables users to read and write emails without needing to be at a computer. The pen writes on any surface, captures a message and displays it on the screen. Twisting the top of the pen selects an address and a press of the button sends.

4.2.2 Gesture-Based Control

Gesture-based controls exploit the natural movements of the hand (and sometimes body), making the hand a tool instead of the means by which another tool is manipulated. Hand movements are critical non-speech components to natural interpersonal communication and offer great freedom in developing more natural communication between persons and computers. Useful hand gestures can range from simple actions such as pointing to complex manipulative ones that express feelings or move objects. Gesture-based systems even include more complex symbolic applications such as American Sign Language (ASL). Proper recognition of sign language could play a vital role in increasing communications between physically impaired users.

Note that although the terms are used interchangeably, "gesture" formally refers to dynamic hand or body signs, while "posture" refers to static positions or poses.

Section 4.2.2.1 provides a brief overview of the technologies used to capture gestures. More specific examples of gesture-based systems and the research surrounding their use comprise Section 4.2.2.2. Section 4.2.2.3 discusses more recent issues of two-handed gesturing as we look forward to complete, seamless HCI.

4.2.2.1   Methods to capture gesture

There are a number of techniques for capturing and interpreting gestures. To exploit gestures for control, there must be a means by which they can be captured or measured and then interpreted by computers. Calhoun and McMillan (2001) explain that gestures can be measured in a variety of ways using several types of hardware devices. Approaches to gesture capture/recognition fall into four major categories:

- Gloves
- Trackers
- Video systems
- Contact devices

Glove-based devices are the most common. They mechanically measure angles, relative positions, etc., while being worn by the user. Glove-based systems are sometimes cumbersome, although very useful in some virtual reality/simulation environments (Pavlovic, 1997). Various measuring technologies have been used within the glove platform including fiber optics, resistance variation, or accelerometers. The CyberGlove by Virtual Technologies, Inc., uses strain gauges to measure finger, thumb, and palm relationships. Repeatability, precision, and reliability are the common problems encountered with glove-based systems from a technical perspective.

Contact devices include classical ones like mice, trackballs, light pens, and touch screens. They typically move in 2-D and operate via direct translation to 2-D screen space. These devices, while commonly used as simple manual input devices discussed in the previous section, can be used in more complex applications like gesture or handwriting recognition. Because one hand is generally dedicated to the contact device, these devices are somewhat limited/constrained.

Tracker devices enable the measurement of the position of an object in space in real time – such as the head, hand, or arm. Traditionally, tracking systems fall into four categories: Mechanical, Electromagnetic, Ultrasonic, and Optical. These four categories are described in Table 4-1 below. Newer tracking systems are now featuring inclinometers, gyroscopes, compasses, and accelerometers. Telemetry can be used to avoid range-limiting connections. These systems are typically less expensive but also less accurate.

Table 4-1: Gesture Tracking Technologies

Tracker Type	General Function	Advantages	Disadvantages
Mechanical	Connect tracked object to potentiometers using rods and/or cables	• High update rates • Low latencies • Inexpensive	• Small range • Impairs free movement
Electromagnetic	Transmitter sends out electromagnetic signal that is detected by a sensor attached to the tracked object. Signal detection varies according to position relative to the transmitter.	• Most precise of non-contact techniques • Large operating range	• Moderately expensive • Metallic objects and other EMI within tracking region • Slight limitation on movement because of connection to sensor
Ultrasonic	Ultrasonic pulses used to determine distances	• Work in metallic environments • Less expensive than EM	• Require direct line-of-sight between emitter and sensor • Higher latency
Optical	Utilize light emitting diodes or reflective dots located on the tracked object – cameras track position and measure locations by correlating to 3-D coordinates.	• Large range of motion/movement	• Require direct line-of-sight between emitter and sensor

Visual interpretation of hand gestures offers a contactless approach and helps to avoid some of the problems found with gloves and trackers, achieving the desired ease and naturalness for improved HCI. Visual interpretation has been approached in many ways - some focussing on hand tracking and hand posture while others attempt to classify hand poses. Most studies are done within a context of a particular application – such as using a finger as a pointer to control television or interpretation of sign language. Most work has been done on recognition of static hand gestures or postures. There is growing interest in dynamic characteristics of gestures because the movements of the hands often contain as much, if not more, information as do the postures (Pavlovic, 1997). Because vision-based systems use cameras to monitor silhouettes of hands or bodies, there are common limitations:

- Limited resolution of the cameras can make it difficult to recognize small elements such as fingers
- Range of movement while using these systems is restricted to the field of view of the cameras
- Fingers and hands can also be obstructed by other body pieces of equipment, making them reliant on clear lines-of-sight
- These applications are usually limited to the situation for they are specifically designed

The optimal choice depends on the environment in which it is used and the tasks it is used to perform. Hand position trackers, for example, work well in benign environments but become unusable under acceleration and vibration found in vehicles and aircraft. Contact devices require space; gloves need to be integrated with existing equipment/clothing and may decrease dexterity needed to operate other equipment. Lags in system performance may be problematic when multi-tasking when the operator cannot wait for the system to respond (Calhoun and McMillan 2001).

Any device involving the hands must address tactile and kinesthetic feedback mechanisms, which are still being developed. The feedback thus provided plays a critical role particularly in object manipulation and robotic applications, as discussed earlier in section 3.4.

4.2.2.2 Gesture Application

Hand gestures are basically used for either manipulation or communication. Manipulative roles are more straightforward and take the spotlight in most HCI applications. Most hand gesture systems are manipulators of virtual objects including 2D and 3D objects, control panels, and robotics. The communicative aspects of hand gestures are subtle and often support speech interfaces. Communicative gestures can affirm and complement the meaning of speech messages. In fact, hand gestures are particularly well suited for multimodal applications (Sharma, 1998). Two examples of communicative systems are discussed below, while several gesture-speech combinations are discussed in more detail in Section 4.6.

Researchers at the University of New South Wales, Australia, are aiming to develop a set of gloves capable of translating Australian sign language in an effort to ease communication between deaf and mute people (AP, 2002). The gloves would be connected to a computer that can measure the movement of the wearer's hand and distinguish between different signs. It then translates the signs into written text on a monitor. A recent trial proved the system to correctly identify signs 95% of the time. Future embodiments could incorporate more wearable displays. Eventually, the project hopes to have the signs translated into spoken words to communicate person-to-person.

The Army's "Digital MP" program features an electronic glove that allows military police to communicate silently using hand signals while separated by woods, buildings, or darkness. Hand signals are designated for certain situations/commands such as "Suspect is armed." Bend sensors in each finger and in the wrist, pressure sensors in the index and middle fingertips, and 2-degree tilt sensors allow hand gestures to be measured and translated into text in the fellow MP's eyeglass display ("Digital MPs," 2001).

4.2.2.3 Two-handed Gesturing

Humans often use both hands, particularly for communicative purposes. Until recently, the single-handed approach was almost inevitable (Pavlovic, 1997). Interpretation and

application of two-handed gestures must deal with several additional challenges such as occlusion of hands/fingers and being able to determine index distinction (left vs. right hand).

Manipulative interfaces using pointing devices could be more efficient with the addition of a second pointing device. Chatty performed a study featuring a two-handed device designed for air traffic controllers. Typically, graphical interfaces are controlled using a pointing device of some kind, manipulated with the dominant hand. Interactions with menus, button, etc., may be improved using two hands since many "real world" manipulations require two hands. Air traffic control environments are composed of maps and a number of symbols depicting waypoints, aircraft position, and speed.

Two-handed systems can be implemented in several ways. A simple way to extend one-handed interfaces is to add a second pointing device that can be used in the same way as the first. Selection and operation times are decreased – one hand can select tools, for example, while the other hand is ready to manipulate the object. A second way is to combine the actions of two pointing devices. Non-dominant hand operations are typically used to hold objects while being manipulated or to add strength.

4.3 Gaze-based Control

Similar to harnessing gestures, machines can also track users' gaze. Harnessing the direction of a user's gaze is a natural and efficient control interface because human beings naturally look at objects they want to manipulate. This can be achieved using head- or eye-based tracking. Using the head as the foundation for gaze-based control relies on the critical assumption that the operator is looking in the general direction that the head is pointing. Recent advances have made eye tracking more available. Eye tracking almost always contains some head tracking as well (Calhoun, 2001).

One common application for using gaze as a control modality is for object selection. Users can select objects simply by looking at them for a set amount of time (usually between 30ms and 250ms) and the computer matches line-of-sight and dwell time. Proper thresholds can be difficult to establish as short dwell times result in inadvertent selection and long dwell times eliminate some of the advantage of gaze-based control. The solution to keeping dwell times to a minimum while avoiding inadvertent selection is often the use of a "consent response." Items are highlighted using gaze but selected using a secondary action such as a button (Calhoun, 2001).

4.3.1 Head Gaze

While "gaze" is truly defined by line of sight, most rudimentary gaze-based systems track the direction of the head. The advantage of using the head is that humans can hold their head in position with relatively good accuracy and stability. For most applications, however, people with full physical capacity do not prefer head-based systems because the movements are frequent and unnatural.

According to Adams and Bentz (1995), head-based control is well suited for non-command applications – becoming more of an input device that changes according to gaze. For example, a virtual environment that changes the scene depending upon direction of head or a helmet-mounted display that displays critical information based on where the pilot/operator is looking. Two examples of command applications where head-based gaze input does work well are aviation and rehabilitation. These two situations have short duration and are inappropriate for traditional interfaces. The aviation community features helmet-mounted sights that aim weapon systems and lock with the head. Most weapon systems are only activated with secondary manual consent. In rehabilitation settings, head-based systems allow some physically impaired operators to interact with computers.

4.3.2 Eye Gaze

Eye tracking is a more direct and accurate measure of gaze. The eye may be advantageous to use for tracking and control because of its speed, accuracy, and stability. This can increase the speed of control operations. Ware and Mikaelian found that object selection and cursor positioning tasks were performed approximately twice as fast with eye tracker as with a traditional mouse (Calhoun, 2001). The downside to eye tracking as means of control is that eye movements are largely subconscious and it is rather difficult to control eye movements in precise ways. For this reason, eye line-of-sight is best used in conjunction with other interface modalities to command activity.

4.4 Speech-Based Control

In the search for more natural means of communicating with machines, speech generally holds the promise of being the most natural and easy. Word processing programs already accept speech interfaces where the user simply speaks commands rather than typing them. A decade ago, Apple included early speech recognition software with the Macintosh systems. They could recognize only a few dozen commands and operated at such a slow pace that they offered no real benefit in terms of productivity. But, DARPA put millions of dollars into automated speech transcription and by 1996 high-end PCs were running Dragon System's *NaturallySpeaking* software. IBM developed *ViaVoice* as a competitor and both now enable the user to select menu options, push buttons, check email, open folders, browse the web, and move the cursor around the screen (Gibbs, 2002). IBM Research now has over 100 researchers working on speech technologies and a similar number working on natural-language understanding.

Speech-based interface facilitate more complex operations, free the hands for other tasks/operations (or free them from action all together), reduce physical space taken up by control device, create access for those with physical/motor impairment, and eliminate the need to see menu items in screen to select. One particular benefit of speech is that speech is usually accompanied by other visible actions such as lip movements, which can be exploited to help clarify the recognition of words (Sharma, 1998). It has been demonstrated that recognition rates for speech can be improved by using visual sensing to analyze lip motions simultaneously. Speech alone can be difficult to recognize with

high accuracy, especially when trying to develop systems that understand free, open language commands versus those that have a preset number of established commands. The advantages of multiple modalities are explored further in Section 4.6. The following two sections (4.4.1 and 4.4.2) review speech as a browsing interface and a more complicated speech-based project at MIT called "Oxygen."

4.4.1 Browsing with Speech

There have been several attempts at adding speech-browsing capability to web-based applications. The main advantage of speech for browsing-type applications is that speech commands are not limited to small screen areas. The user may choose from more options than can be seen on-screen; links do not need visible hyperlinks. This could be very helpful for pages that are frequently visited and already known, but is somewhat reliant on a site being arranged in a logical and predictive manner.

Borges *et al* (1999) investigated speech as a viable means of browsing the web. Their study aimed to determine the effectiveness of speech as a browsing modality but also the preferences of the users. While participants were free to interact naturally with the computer, their communication was limited to very simple phrases, often consisting of only one or two words. It appeared that they were behaving in accordance with what they thought the language of the application was rather than using their own "natural" language. Interestingly, more than half of the issued commands were for accessing pages as most of the user interaction was concentrated on moving from one page to another. Other researchers have found similar results where users simplify language structure when interacting with computers. Borges *et al* also report that previous studies indicate users prefer speech interfaces when small vocabularies are involved. This is somewhat contradictory to the idea that we need to achieve natural conversations with computers.

Earlier work at the Center for Spoken Language Understanding at the Oregon Graduate Institute resulted in Spoken Language Access to Multimedia (SLAM) - a graphical user interface for browsing the web with speech capabilities. One of the conclusions made by the SLAM team is that speech is ideal for multimodal systems because of the way it complements the typical mouse/pointer-based systems (Borges, 1999). The SLAM work introduces the idea of using speech as one modality alongside other complementary modalities. Most often, we see speech used in conjunction with manual controls (keyboards, mice/pointers), gesture controls, and facial recognition. When compared to pointer-based browsing, speech-based interaction typically resulted in lower average time. Time to complete tasks, however, is too highly dependent upon the technology used, the tasks to be completed, and a variety of external factors. Studies demonstrate that users actually preferred speech interaction to manual (keyboards) even though interaction times were greater. See Section 4.6 for more detailed discussion of multimodal systems.

4.4.2 Project Oxygen

A more complex speech-based project under development is MIT's "Oxygen" project. Oxygen is aimed at creating the kind of interaction between computers and people that we typically see in the movies, giving computers the capacity for human-like interaction. The ultimate goal is for machines to recognize users (via facial recognition), enable the user to ask questions in casual conversational language, and ask other machines for help without being told. Building in multiple modalities of interaction, Oxygen has just gotten to the point of recognizing pointing gestures in addition to speech recognition. Pointing at a screen can make a dot appear; arm movements determine corresponding dot movements. Oxygen currently has a voice-controlled office where spoken commands are used to start presentations, open blinds, etc. Several of the misrecognition errors discussed earlier take place at MIT as the computer still has trouble distinguishing commands and noise. Non-commands result in inadvertent actions. While significant advances have been made in speech recognition, speech-based interfaces are not flawless. Misrecognition errors remain a barrier. Three basic classes of error include:

- Substitution error: recognizer substitutes a word in its vocabulary for a word that was actually spoken
- Rejection error: recognizer cannot form hypothesis about utterance and rejects the command all together
- Insertion error: recognizer cannot recognize because background noise or non-verbal noise disrupts speech recognition

Oxygen is working to refine speech recognition with Victor Zue. Current Oxygen technology is limited to a number of spoken commands. User must train the system to recognize speech patterns, and commands must be issued precisely in order for the system to function properly. To help experiment with some of the underlying issues, the team is running a speech-activated telephone system that provides weather and traffic information that allows more open queries. A user can ask, "What's the temperature?" or "How hot is it?" and the system will answer either way. The telephone system can also recognize multiple languages and operate without a "training period."

4.5 Neural (Brain-Actuated) Control

Neural control refers to harnessing the electrical activity of the brain to control devices. The hands-free aspects of this approach make it an attractive option particularly for situations where hands are needed for other tasks or where hands are not functional because of paralysis or other physical impairment. It is also non-fatiguing. The notion of controlling a device simply by thinking is seen as the ultimate in intuitive control (McMillan, 2001). It is envisioned that people may someday no only control typical computer interfaces (such as a mouse moving a cursor) but also control wheelchairs or robotic arms to replace the lost functions of natural limbs. This far-reaching goal is referred to as neuroprosthetics.

The brain's electrical activity can be measured by invasive or noninvasive means. The more invasive approach is to implant arrays of microwires into specific areas of the brain responsible for the action/control needed. While invasive, it is a more robust technique as specific neurons can be targeted, which allows more precise control. The noninvasive approach measures electro-encephalographic (EEG) signals associated with brain activity at the surface of the skull, and is useful for a different type of interaction. Because EEG represents the average electrical activity of broad populations of neurons, they cannot be used directly for limb prosthetics or precise control of movements. EEG signals are more effective for simple interactions such as selecting letters on a computer screen. Section 4.5.1 describes a system based on EEG control. Section 4.5.2 discusses the more invasive microwire array technique being researched at universities around the world.

4.5.1 EEG Systems

Brain Actuated Technologies, Inc., has created a multimodal control system including brain-actuated control. Their Cyberlink System combines eye-movement, facial muscle, and brain wave bio-potentials detected at the user's forehead to generate computer inputs that can be used for a variety of tasks (http://www.brainfingers.com/index.html). The forehead offers a rich variety of signals that can be reached in relatively non-invasive ways. Three different types (or channels) of control signals are derived from the forehead signals by the Cyberlink Interface and can be used to control vertical and horizontal cursor motion, functions of the left and right mouse buttons, on/off switch control, on/off program commands, and some keyboard commands. In a discrete control study conducted by the United States Air Force at Wright Patterson Air Force Base in Dayton, Ohio, subjects' reaction times to visual stimuli were found to be 15% faster with the Cyberlink EMG button than with a manual button.

Specific facial and eye movement gestures can be discriminated by the Cyberlink software and mapped to separate mouse, keyboard, and program functions. This hands-free mouse enables the user to steer the cursor, change its speed and resolution, perform left and right mouse button functions, and send keyboard characters and character string commands. In a recent study, users were able to use the mouse to position and click the cursor over randomly appearing 32 x 32 pixel (icon-sized) targets in 4 seconds or less.

4.5.2 Implantable Arrays

In the mid-1990's, researchers at Hahnemann University taught a rat in a cage to control a lever with its mind. After a period of "training" or conditioning, rats realized that they no longer needed to press a bar to be rewarded with a drop of water. If they just looked at the bar and imagined its forelimb pressing it, neurons express the firing pattern that is interpreted as motor commands to move the lever. Rats learned that they only had to think through the action of pressing the bar.

Further work with monkeys showed similar success. Using more complex brain structures allows scientists to investigate likelihood of eventual human applications.

Ongoing research at Duke University and MIT investigated a monkey's ability to control robotic arms via brain control. Wearing a cap that fed microwires into its motor cortex, a monkey was able to control two dissimilar robotic arms at the same time. The monkey responded to flashing lights by moving a joystick in the direction of the light while the electrical activity of the neurons were captured and used to move the robotic arms in sync with the monkey's real arm movements. More recent studies have suggested that sensory feedback can allow people to improve the performance of brain-machine interfaces. Experiments are underway to investigate how performance can be improved by providing visual and tactile feedback while using brain control.

There are still hurdles to overcome before brain-machine interfaces are safe, efficient, and reliable options. Surgical electrode implants will always be of medical concern and scientists need to learn more about long-term impact on human brain tissue. Development of lightweight, dense microwire arrays is underway and increases the number of neurons that can simultaneously recorded.

More and more scientists are embracing the vision that brain control devices can help people. Traditional neurological laboratories have begun to pursue neuroprosthetic devices. Preliminary results have appeared at Arizona State University, Brown University, where rhesus macaque monkey was shown to move a cursor around a computer screen. In the distant future, neuroscientists may be able to regenerate injured neurons or program stem cells to take their place. Until then, brain-machine interfaces are a more viable option for restoring motor function (Nicolelis and Chapin, 2002).

4.6   Multimodal Control Systems

As growing attention is placed on the feasibility and utility of individual modalities, a limiting feature of modern interfaces is that they remain largely unimodal; i.e., they rely on one mode of interactions such as a mouse movement, a key press, or speech input. Although one interaction modality may be adequate in many cases, there may be circumstances where improved task performance can be achieved through the use of multiple control modalities. In manipulating a 3D object, for example, a user might have to select an object with a mouse and then use the same mouse to select the correct control panel menu to change the objects' color. It would be much easier and more natural if the user could point at the object and say "make it green." Almost any natural communication among humans involves multiple, concurrent modes of communicating. We speak about, point at, and look at objects. We hear tone of voice and look at expressions and body movements to deduce clues about emotions. The ease with which unimodal interaction allows us to interact with computers is often unsatisfactory.

The reasoning behind multimodal interactions were organized by Sharma *et al* (1998) into the following three categories: Practical, Biological, and Mathematical. The following three sections review the practical, biological, and mathematical reasons for wanting multimodal interfaces. Section 4.6.4 then reviews some top-level recommendations for using speech and gesture, as they are the most commonly combined modalities found in literature.

4.6.1 Practical Reasons

Practical reasons stem from the inherent shortcomings of modern HCI systems that are ineffective, unnatural, and cumbersome. There are six practical reasons for using multiple modalities for human-computer interaction:

1) Users prefer multiple modalities
2) Multiple modalities compliment one another
3) Multiple modalities increase robustness of communication
4) Multiple modalities help ensure more universal access to technology for various user groups
5) Different modalities are better suited for expressing different things
6) Multiple modalities better enable multitasking

Sections 4.6.1.1 through 4.6.1.6 discuss each of these in more detail. Of these practical reasons, most of the literature supports the first three: User Preference, Complementary, and Robustness and Accuracy.

4.6.1.1 User Preference

Several studies have concluded that people prefer to use multiple modalities, particularly for virtual object manipulation. Hauptmann and McAvinney (1993) found that 71% of test subjects preferred to use both speech and hands to manipulate virtual objects. The more spatial the task, the more they prefer multimodal interfaces. Oviatt *et al* (1997) has shown that 95% of the subjects in a map manipulation task tend to use gestures with speech. Pavlovic *et al* (1997) also showed speech and gesture to be a preferred combination in controlling virtual environments. Psychological studies have shown that people prefer to use hand gestures in combination with speech in virtual environments because they allow good control without training or special apparatus.

The Hauptmann and McAvinney study asked three user groups to communicate with a computer using either speech alone, gestures alone, or speech and gesture together. The task was to manipulate a 3D cube on the screen. In addition to commonality shown in use of gestures and speech, users showed a preference for using speech and gesture in combination. All subjects completed all tasks. This study provides a foundation for understanding how people use speech and gestures to communicate with computers:

- Speech data showed that small number of total words used with few words spoken at a time. This confirms earlier studies that also found limited vocabularies adequate for communication, despite the interest in developing fully capable natural communications.
- Gesture data indicates that users are not comfortable using single finger gestures – at least for the cube manipulations in this experiment. Users made gestures in all three dimensions, not just in the plane of the screen.
- Users prefer combination of gesture and speech over speech or gesture alone.

- There was a surprising uniformity in the way both speech and gesture were used, indicating there may be an intuitive, common principle in gesture communications. Spatial tasks with gesture manipulation were equally accessible to novices, experts, and proficient graphic manipulators.

4.6.1.2 Robustness and Accuracy

Modern speech recognition systems are still error-prone, especially if used alone. The use of two systems together can reduce the uncertainty or ambiguity, lessening the restrictions needed for accurate interactions and reducing the complexity of creating a "natural" HCI with one mode of interaction. Spoken words can affirm gestures, and gestures can clarify noisy or ambiguous speech. Introducing gaze with speech and gesture can make the system even more robust. Oviatt found that multimodal input (speech and pen) produced a 36% reduction in task errors and 23% fewer spoken words. Each modality can also be used to correct ambiguities in the other. For example, a spoken command "Create blue box here" can be made less ambiguous if it is known where the user is pointing.

Rather than having to develop complicated gesture-only systems, designers can simplify the process by employing speech and gesture together in ways that do not require such strict interpretation. There were large amounts of work completed between 1993 and 1997 regarding integration of multiple modes. Most studies, while searching for how to combine or integrate modes, at least point toward resolution of ambiguity as an advantage. Vo and Waibel (1997) studied the integration of speech and lip-reading - recognition rates of bimodal systems were always better than or equal to either of the unimodal rates. The combination of speech and lip-reading resulted in 9% reduction in error over "clean" speech-only and 29% error reduction over "noisy" speech-only conditions (even though lip-reading-only had a mere 12% recognition accuracy). Nakagawa *et al* experimented with speech and haptics and found advantages to using speech and haptics together. Again, bimodal condition had better recognition rates because haptics served to enhance available information when speech was poor or ambiguous (Mills and Alty, 1998).

4.6.1.3 Complementary

Multiple modalities complement one another. Voice and gesture together create an input more powerful than either one alone. The strength of one makes up for the lacking of the other. Cohen showed, for example, that gestures are ideal for direct object manipulation while speech/natural language is best suited for descriptive tasks. The strengths of one make up for the weakness of another (Cohen, 1989). It is thought that speech and gesture make ideal complements to one another for communicating.

Humans constantly send a mixture of complementary and redundant information, allowing them to achieve high success rate in communicating intentions to one another. This concept ranges back to 1980 when MIT performed the "Put-that-there" program (Mills and Alty, 1998). The "Put-That-There" work by the Architecture Machine Group

is the seminal multimodal graphical interface. Wrist-mounted magnetic position sensing devices measured hand positions, and users employed speech and gesture (or a combination of them both) to add, delete, and move graphical objects shown on a projection wall. Done nearly 20 years ago, it is still impressive work in the multimodal area because they discovered that integrating speech and gesture with contextual understanding allowed neither system to have to perform perfectly as long as complimentary modalities converged on the intended meaning together.

4.6.1.4 Universal Access

Universal access (for physically/mentally handicapped) is promoted by use of hand gestures for ASL, eyes tracking, speech recognition, and EEG-based control. All help the physically challenged gain access to an otherwise restricted world of information.

4.6.1.5 Ease of Expression

Typical computer interactions can be thought of in terms of "Ease vs. Expressive" tradeoffs. Mice are easy but lack an expansive vocabulary for expressing; keyboards are not easy but maximize expressiveness. Multiple modalities tend to overcome both aspects of interaction.

4.6.1.6 Multitasking

A person's ability to perform multiple tasks is affected by whether the tasks use the same or different sensory modes. In multimodal interfaces users can perform visual/spatial tasks at the same time as giving verbal commands. Users of a CAD program were able to be more productive using speech and manual control – remaining visually concentrated on the screen while using speech commands.

4.6.2 Biological Reasons

Nature is another source of rationale for multiple modality interactions. Humans as well as other animals integrate multiple senses all the time in natural communication. Studies of the brain show that different senses are initially segregated at the neural level. Reaching the brain, they converge at a certain location (superior colliculus) and are further processed. About 75% of the neurons leaving the superior colliculus are multisensory (Murphy, 1996).

4.6.3 Mathematical Reasons

The field of sensory data fusion points us in the direction of multimodal integration because of its thrust toward target detection. The goal is to find optimal ways to integrate different sensory data that produce the best detection rates. Statistical analysis explains that using a single sensory input may not be adequate for the basis of decision making – redundant data improves the ability to understand (Murphy, 1996).

4.6.4 Recommendations

As can be seen, there are powerful effects achieved by combining speech and gesture recognition. Current speech and gesture technologies make multimodal interfaces with combined modalities easily available as well. Having this multimodal capability, however, does not mean voice and gesture should be added to every package. Intuitive interfaces require thought and planning to utilize the strengths of both modalities. There are several recommendations for their use in interface development with the overlapping theme being to use multimodal interfaces to develop contextual understanding that reduces ambiguity. Table 4-2 provides guidelines for the use of speech or gesture modalities. (Billinghurst, 2002).

Table 4-2: Guidelines for Selection of Speech or Gesture Modalities

Speech	Gesture
Need for special acoustically distinct command vocabulary	Hand tension should signify start of command
Provide constant feedback about recognizer activity	Commands should be fast, incremental, and reversible
Separate speech from graphics as much as possible	Natural gestures should be favored for ease of learning

4.7 Summary

In this chapter we have presented an overview of the different control modalities that can be used for controlling or manipulating a system via a human/machine interface: manual, gaze, speech, and neural. Each one has strengths and weaknesses with respect to different types of tasks. While manual control technologies are the most mature, gaze and speech-based control are becoming increasingly viable. Neural control is a relatively new technique that offers some promise. It may prove to be an effective means by which disabled or paralyzed individuals can interact with the technology around them. The tables below outline the applicability and benefits of the control modalities discussed in this chapter.

Table 4-3: Applicability of Manual Control

Technology or Methodology	Applicability

Simple	• Manual controls are generally suited for controlling spatial elements and are often paired with visual presentations (see stimulus, manipulate with hands)
	• Manual controls also used for verbal information such as text entry, but not always ideal (see Table 4-5 for Speech)
	• Simple manual applications include:
	  • Switches for on/off functions
	  • Keyboards for fast and accurate text entry or edit (quieter than speech entry if covert is needed)
	  • Mice best for selecting and moving graphical items
	  • Pens and tablets used for drawing or handwriting because of finer muscle controls of fingers
	• Feedback needed with manual controls to indicate status (feedback may be visual, auditory, tactile, etc)

Gesture	• Gesture controls have two major applications: Manipulation and Communication.

	- Spatial manipulation of robotic arms and grasping virtual objects - Gestures aid communication by confirming meaning of speech (the most popular multimodal combination found in our literature review) - Gesture is primary communicative channel in sign-language recognition or other gesture codes • Sign-language applications may be coupled with visual textual display to show the signed message, eliminating some time and cognitive load needed for interpretation

Table 4-4: Applicability of Gaze-Based Control

Technology or Methodology	Applicability
Head	• Head-based gaze controls are best for infrequent, non-command applications because long or repeated movements of head can become annoying, tiresome, and distracting from other tasks • Best used for tasks where the control is closely associated with gaze such as moving or selecting items; the head can be held in position with good accuracy and stability • Also good to consider gaze control when physical impairment limits other control modalities
Eye	• Eye-based gaze is more applicable for frequent controls because the eye can better handle frequent, repetitive movements • Eye offer accuracy and stability of head, along with speed

Table 4-5: Applicability of Speech-Based Control

Technology or Methodology	Applicability
Speech-based Browsing	• Verbal commands and text entry can be "more natural" with speech compared to typing or handwriting recognition • Speech control allows user to activate options/functions without seeing and selecting on screen • User preference studies have shown speech to be preferred over manual interfaces for browsing and that small, simple vocabularies to be preferred
Commands	• Speech can be used for just about any operating commands where background noise is not a cause for error (should improve as voice recognition technology improves) – open doors, turn on lights, change the channel, etc. • Speech in general frees hands to be used in other concurrent tasks

Table 4-6: Applicability of Neural Control

Technology or Methodology	Applicability
EEG systems	• Surface EEG systems are most effective for simple spatial interactions such as moving cursors and selecting items or letters on the screen (because they measure broad activity of the brain) • Used where hands are needed for other tasks or if physical impairment limits use of limbs • Suspect neural controls to be highly impacted by multi-tasking or shared attention tasks • Can be more robust with multimodal systems such as muscle-based controls, gaze tracking, etc.
Implantable Array	• Implantable arrays offer more specific measurements of brain activity • Added precision allows more complex control of prosthetics or other controls • Most basic research focuses on activating manual controls without having to be manual (such as pulling a lever by simply thinking about pulling the lever)

HMIs have traditionally centered on the use of a single control modality, which is rather unlike normal human communication. Almost any natural communication involves multiple modalities – we point at objects, speak to people, and deduce meaning from movements. There are numerous reasons to move toward multimodal interfaces to achieve improved interactions. Whether used singly or together, the task, environment, and type of processing should guide the selection of controls needed to operate a system.

5. Analysis of Soldier Needs

5.1 Introduction

Information display and control components associated with soldier systems are typically developed according to the way the information has historically been provided, or on the type of technology that is currently available. Developing information display systems for soldiers by first examining what their information needs are and then determining what the best modality or combination of modalities would be to present that information, would lead to more robust, usable, and effective soldier systems. The type of information being displayed, the action being performed, and the information processing to be done, should drive the development of information display technologies for soldier systems, not vice versa.

In this chapter we examine how the principles and design methodologies discussed in the preceding three chapters can be used to support task-oriented identification of HMI modalities for soldier systems. Section 5.2 identifies the relevant soldier needs, based on work conducted under the Scorpion and OFW programs to characterize soldier tasks and activities. In section 5.3 we describe the process by which these tasks were classified according to the type of cognitive processing they entail (following the principles discussed earlier in Chapter 2). Finally, section 5.4 presents some preliminary thoughts on which HMI modalities offer the potential to best support human performance in supporting these tasks. This analysis is based on making a connection between the nature of the cognitive processing associated with a task and the findings from the literature on the merits and applicability of various input/output modalities, as discussed earlier in chapters 3 and 4.

5.2 Identification of Soldier Needs

As a sample top-level framework for this type of information-driven system development, we considered the OFW soldier needs as an example. Starting with the complete list of OFW needs (Preliminary Master List of OFW Needs and Reconciliation with Scorpion Needs ~ Results as of Workshop I, Last Revised 9/4/02), we downselected the list to include only those needs with a significant cognitive component. Needs that are purely physical in nature were not considered in this exercise. Table 5-1 presents the resulting subset of eighteen needs from the OFW soldier needs list together with their definitions.

Table 5-1: OFW Need Definitions

OFW Need #	OFW Need Title	Preliminary OFW Need Definition
N1	Mission planning and rehearsal	To prepare and rehearse for any mission (including special team rehearsals) exploiting embedded virtual and constructive means.
N2	Intelligence Collection	To collect intelligence from a variety of sources (i.e., onboard and remote sensors, HUMINT, etc.).
N3	Intelligence Dissemination	To securely distribute/ receive relevant intelligence information in near-real-time, down to the lowest level, and with a low probability of intercept.

OFW Need #	OFW Need Title	Preliminary OFW Need Definition
N4	Position/location/ Tracking	To determine location of self and others, within one meter, to include in, out, and around structures and all terrain conditions.
N5	Mapping	To represent all types of terrain and man-made structures, and to overlay relevant information.
N6	Navigation	To move from one location to another, safely.
N7	ID friendly, enemy, non-combatants	To identify combatants under all conditions.
N13	Enhanced vision; "see thru/past" obstacles	To allow individual to see under all conditions, including transition between light and dark, beyond-line-of-sight, and to be able to know what is on the other side of an opaque obstacle. (*New definition per instructions from Workshop I*)
N14	Detect/avoid hazardous areas	To detect, defeat, and/or bypass any hazardous area.
N15	Situational Understanding	To comprehend, the locations and interactions among persons, terrain, and objects under all conditions, appropriate for the level of the individual.
N16	Target designation	To correctly assign a target under all conditions.
N17	Target detection and recognition	To detect, recognize, and identify a target, under all conditions.
N20	Synchronization of fires	To interactively coordinate, within the unit of action, the plans and execution of fires across the company-level battlespace.
N21	Direct engagement	To directly engage persons or other targets through a variety of lethal means and methods.
N24	Target hand-off	To hand-off a target, verify correctness of same, prioritize and coordinate with appropriate Objective Force assets.
N25	Communications	To send and receive secure and non-secure voice, data, imagery, non-line-of-sight and non-RF; with hands free, selectable LPI, LPD, and anti-jam operation.
N26	Information management	To provide relevant information to the right person at the right time in a usable form to facilitate situational understanding and decision making.
N35	Sustain and Enhance Individual Performance	To sustain and enhance physical and mental performance.

5.3   Classification of Soldier Tasks

To assess what modality or combination of modalities might be best suited to each OFW soldier need, it was necessary to classify needs according the types of information processing (either verbal or spatial) associated with each need. More details about information processing are available in Chapter 2. Characteristics of spatial and verbal information processing tasks are presented below in Tables 5-2 and 5-3.

Table 5-2: Characteristics of Spatial Tasks

Characteristic	Definitions
Judgement concerning axes or translation or rotation	Visualization of space or items in space, visualization of 3-D objects or environments, maps, etc.
Motion perception and tracking	Perceive and track the motion of other moving entities in the environment
Interpolation or extrapolation of continuous functions	Decisions or perception related to movement, acceleration or deceleration, or trends in movement. Perception of movement of self relative to others.
Localization of self and/or others	Developing mental model of own location and that of others. "Others" includes physical entities, structures, landmarks, etc. as well as other people.

Table 5-3: Characteristics of Verbal Tasks

Characteristics	Definitions
Use of language	Communication of written or spoken information.
Use of arbitrary symbolic coding	Reasoning involving general symbols, icons, and abstract representations of real-world information.
Mental arithmetic	Calculations of distance, time, ordering, priority, etc.
Rehearsal	Review of steps or actions to be taken, includes checking against a plan.

Soldier needs are a combination of tasks and information requirements that are associated with meeting each need. In our assessment we considered the information required to meet each need, and not the method or technology by which the need is currently met. For example, item N5 in Table 5-1 requires that the soldier be provided with information about the location of important landmarks, natural features, and access routes (roads, trails, etc.). We did not consider how that information is currently provided (i.e., with paper maps or digitized displays). Rather, we only considered the nature of the required information and therefore the type of information processing involved in tasks required to meet the need for mapping.

Table 5-4 presents each of the OFW needs we considered and how they were categorized in terms of information processing. The type of information processing associated with meeting each soldier need is what is then used to determine which modality or combination of modalities should be employed when developing information presentation technologies aimed at satisfying that particular soldier need.

Table 5-4: OFW Needs – Classification of Information Processing Required

	OFW Need	Spatial Information Processing				Verbal Information Processing				Comments
		judgement concerning axes or translation or rotation	motion perception and tracking	interpolation or extrapolation of continuous functions	localization of self and/or others	use of language	use of arbitrary symbolic coding	mental arithmetic	rehearsal	
N1	Mission planning and rehearsal			X	X	X	X		X	Includes knowing where yourself and others in your unit are.
N2	Intelligence Collection				X	X	X	X		N2 & N3 have the same information needs. Includes knowing where you are and where other things are.
N3	Intelligence Dissemination				X	X	X	X	X	N2 & N3 have the same information needs. Includes knowing where you are and where other things are.
N4	Position/location/Tracking	X	X	X	X		X			Covers static location of people and things and their movements.
N5	Mapping	X			X	X	X	X		Includes the creation and utilization of maps, mapping static landmarks, terrain, etc. Does not include N2 and N3.
N6	Navigation	X		X	X	X		X		This pertains to the individual's own movement – not tracking something or someone else moving
N7	ID friendly, enemy, non-combatants					X	X			This is purely identification, no location or movement judgements.
N13	Enhanced vision; "see thru/past" obstacles	X	X	X	X					Interpolation or extrapolation of continuous functions applies here because the need definition includes things that are moving. The need specifies enhanced "vision" – therefore, no verbal component. If the need were stated as enhanced "knowledge" (i.e. knowledge of what's in a room, without necessarily having to "see" it) the task identifiers would be different. Knowledge is not limited to vision, necessarily.
N14	Detect/avoid hazardous areas	X	X	X	X	X	X			Includes detection of motion.
N15	Situational Understanding	X	X	X	X	X	X	X		
N16	Target designation					X	X			Assign a target to someone.
N17	Target detection and recognition	X	X	X	X	X	X	X		Includes detect a target and recognize what it is, only. (N7 covers IFF decisions and then N16 covers designating a recognized entity as a target).
N20	Synchronization of fires	X	X	X	X	X	X	X	X	
N21	Direct engagement	X	X	X	X					Aim and shoot – spatial only
N24	Target hand-off	X			X	X	X	X		Prioritization of targets requires mental arithmetic.
N25	Communications					X	X			
N26	Information management					X	X	X	X	Playback of previously developed plans = rehearsal.
N35	Sustain and Enhance Individual Performance			X		X	X	X		Physiological Status Monitoring, for example, requires information presentation.

Most needs do not consist exclusively of spatial processing or verbal processing; rather, they include components of both. As a result, they fall somewhere along a continuum whose endpoints are represented verbal and spatial tasks. Of the 18 OFW needs we considered, we found that only six of these eighteen involved entirely verbal or entirely spatial processing:

N7   ID friendly, enemy, non-combatants (verbal)

N13   Enhanced vision; "see thru/past" obstacles (spatial)

N16   Target designation (verbal)

N21   Direct engagement (spatial)

N25   Communications (verbal)

N26   Information management (verbal)

The remaining twelve needs were categorized as requiring a combination of both verbal and spatial processing.

5.4   Candidate Modalities to Support Soldier Needs

In this chapter we have classified broad soldier needs according to the types of information processing they entail. Earlier in Chapter 2 we outlined the models of cognitive processing that explain how tasks and associated information loads can be classified according to the cognitive central processing they require. Chapters 3 and 4 reviewed various modalities of information presentation and control, discussing each in terms of how well it is suited for various information applications. Table 5-5 illustrates how this information can be synthesized to connect human tasks with candidate HMI modalities. Tasks were selected to include examples of those that were classified in Table 5-4 as being spatial, verbal, or both. Exact specification requires detailed consideration of tasks, context, and concurrent activity.

Table 5-5: Candidate Modalities

OFW Need #	OFW Need Title	Processing Classification	Candidate Modalities
N5, N6	Mapping, Navigation	Spatial and verbal	Combination of visual presentation with haptic feedback and/or 3D auditory cues to indicate heading, location, distance, terrain, etc. Manual or speech-based landmark manipulations (assuming electronic map), labeling, etc.
N16	Target designation	Verbal	Audio speech and/or gesture
N21	Direct engagement	Spatial	Eye- or head-based gaze with manual trigger

Admittedly, nothing can be definitive at this level of consideration. Further investigation into these tasks and use scenarios is required for proper specification. In general, the following generalizations are likely to hold true:

- Presentation of spatial information will remain primarily visual with augmentation by spatialized (3D) audio and directional haptic cues

- Manual controls match well with visual presentations for manipulation of spatial elements

- Simple auditory presentation can be used in almost any situation to gain attention or cue alert

- Gesture can be used alone or with speech to enhance communications

- Speech recognition technologies will continue to improve and be viable for browsing and other verbal software applications (limited in covert operations)...

- Gaze-based control can be used for object selection and tasks such as aiming. It is best suited for situations requiring object dwell times that are neither very short (making repeatability and accuracy of interpretation difficult) nor too long (defeating the benefits of using gaze in the first place).

- Neural controls are unlikely to be viable for real-world applications in the near term. They hold considerable promise for enabling disabled people to interact with computing technologies, but at this time they are not sufficiently robust for field use.

- Olfactory cues are effective for (but remain limited to) emergency/notification scenarios where only presence of smell needs to be determined (e.g., detection of natural gas leaks).

- Multimodal controls offer the potential for natural communications between human and machine

- It is critical to consider tasks and modalities as a system within their context of use

6. References

Abouchacra, Kim S., Breitenbach, Mermagen, and Letowski (2001). "Binaural Helmet; Improving Speech Recognition in Noise with Spatialized Sound." *Human Factors, 43(4)*, pp. 584-594.

Adams, K.D. and Bentz, B.D. (1995). "Headpointing: The latest and greatest." *Proceedings of the RESNA 18th Annual Conference,* Vancouver, Canada, pp.446-448.

Adams, K.D. (2001). "Input devices and controls: Manual Controls." *International Encyclopedia of Ergonomics and Human Factors. Vol. 1,* pp. 816-831.

"Australian scientist developing talking gloves for the deaf." Associated Press, 8/22/2002 09:34.

Baca, Julie (1998). "Comparing Effects of Navigational Interface Modalities on Speaker Prosodics." http://www.wes.army.mil/ITL/baca98.html

Billighurst, Mark. "Put that where? Voice and gesture at the graphics interface." www.siggraph.org/publications/newslwtter/v32n4/contributions/billinghurst.html

Biggs, J. and M. Srinivasan (2002). "Haptic Interfaces," in *Handbook of Virtual Environments.* K. Stanney (Ed.). London: Lawrence Earlbaum, Inc.

Blackwood, W., Anderson, T., Bennett, C.T., Corson, J., Endsley, M., Hancock, P., Hochberg, J., Hoffman, J., Kruk, R., Mavor, A., Kidd, J., and Prince, C. (1997). "Tactical Display for Soldiers: Human Factors Considerations." National Research Council. National Academy Press, Washington, DC. PB97-138044

Bolia, R.S., D'Angelo, W.R., McKinley, R.L. (1999). "Aurally Aided Visual Search in Three-Dimentional Space," *Human Factors, 41(4),* pp. 664-669.

Borges, J.A., Jiminez, J., and Rodriguez, J. (1999) "Speech Browsing the World Wide Web." *IEEE, 1999.* 0-7803-5731-0

Brainard, R., Irby, T., Fitts, P., & Alluisi, E. (1962). "Some Variables Influencing the Rate of Gain of Information," *Journal of Experimental Psychology, 63,* pp. 105-110.

http://www.brainfingers.com/index.html

Bray, H. "A Fresh Breathe of Oxygen." *www.digitalMass.com*

Calhoun, G.L. (2001). "Gaze-based Control." *International Encycolpedia of Ergonomics and Human Factors, Vol 1,* pp.234-236.

Calhoun, G.L. and McMillan, G.R. (2001). "Gesture-based Control." *International Encycolpedia of Ergonomics and Human Factors, Vol 1,* pp.237-239.

Chatty, S. (1994). "Issues and experiences in Designing Two-Handed Interaction." *Conference Companion, CHI '94.* Boston, MA, 1994, pp.253-254.

"Computer mouse for the blind developed." www.cnn.com/2002/TECH/ptech/09/09/mouse.blind.reut/index.html

Cook, G. "Company builds game plan to help the blind." www.digitalmass.com.

Cohen, P.R., Darlymple, M., Pereira, F., Sullivan, J.W., Gargan, R.A., Schlossberg, J.L., and Tyler, S.W. (1989) "Synergic use of direct manipulation and natural language," *Proceeding Conf. Human Factors in Computing Systems (CHI'89)*, Austin, TX, pp.227-233.

Daugman, John (1997). "Face and Gesture Recognition: Overview." *IEEE Transactions on Pattern Analysis and Machine intelligence*, Vol.19, No.7, July 1997.

Davide, F., Holmberg, M., & Lundstrom, I. (2001). "Virtual Olfactory Interfaces: Electronic Noses and Olfactory Displays," in G. Riva & F. Davide (Eds.), *Communication Through Virtual Technology: Identity Community and Technology in the Internet Age (Volume 1)*, Amsterdam: IOS Press.

"Digital MPs", U.S. Army Soldier & Biological Chemical Command, U.S. Army Soldier Systems Center-Natick Public Affairs Office, January 29, 2001.

Durlach, N., Delhorne, L., Wong, A., Ko, W., Rabinowitz, W., & Hollerbach, J. (1989). "Manual discrimination and identification of length by the finger-span method," *Perception and Psychophysics*, *46*(1), pp. 29-38.

Gibbs, W.W. (2002). "Whatever you say." www.sciam.com. May 13.

Glumm, M.M., Marshak W.P., Branscome, T.A., McWesler, M., Patton, D.J., and Mullins, L.L. (1998). "A Comparison of Soldier Performance Using Current land navigation Equipment with Information Integrated on a HMD." ARL-TR-1604, April 1998.

Glumm, M.M., Branscome T.A., Patton, D.J., Mullins, L.L., Burton, P.A. (1999). "The Effects of an Auditory Versus Visual Presentation of Information on Soldier Performance." ARL-TR-1992, August 1999.

Greenwald, A. (1970). "A double stimulation test of ideomotor theory with implications for selective attention," *Journal of Experimental Psychology*, *84*, pp. 392-398.

Greenwald, A. (1979). "Time-Sharing, Ideomotor Compatibility and Automaticity," *Proceedings of the 23rd Annual Meeting of the Human Factors Society*, Santa Monica, CA.

Hauptmann, A.G. and McAvinney, P. (1993). "Gestures with Speech for Graphic Manipulation." *International Journal of Man-Machine Studies*, vol.38, pp.231-249.

Hedge, A. (2002). "Multitouch technology – improving the ergonomic design of input devices." http://ergo.human.cornell.edu/CUmultitouch.html

Kantowitz, B., & Knight, J. (1976). "Testing Tapping Timesharing: I. Auditory Secondary Task." *Acta Psychologica*, *40*, pp. 343-362.

Krueger, M. (1995). *Olfactory Stimuli in Virtual Reality Medical Training*. Artificial Reality Corporation, Vernon, CT.

Krueger, M. (2002). *Personal communication.*

Lee, J.D., Caven, B., Haake, S., and Brown, T.L. (2001). "Speech-based interaction with in-vehicle computers: The effect of speech-based email on drivers' attention to the roadway." *Human Factors*, Vol. 43, No. 4, pp.631-640.

Loomis, J.M., Golledge, R.G., Klatzky, R.L., Speigle, J.M., and Tietz, J. (1994). "Personal guidance system for the visually impaired." *ASSETS 94, The First Annual ACM Conference on Assistive Technologies*. Oct31-Nov1, 1994, pp85-91.

Manjoo, F. (2001). "Making Senses Out of Games," *Wired Magazine*. Available online at http://www.wired.com/news/culture/0,1284,42417,00.html.

McMillan, G.R. (2001). "Brain and Muscle Signal-based Control." *International Encyclopedia of Ergonomics and Human Factors, Vol 1*, pp.379-381.

Mills, K.M. and Alty, J.L. (1998). "Investigating the Role of Redundancy in Multimodal Input Systems." In M. Waschsmuth & M. Frohlich (Eds.), *Lecture Notes in Artificial Intelligence*. Berlin: Springer-Verlag.

Mulgund, S.S. and Zacharias, G.L. (1996). "A situation-driven adaptive pilot/vehicle interface." *Proceedings of the IEEE Human Interaction with Complex Systems Symposium*, Dayton, OH.

Murphy, R. (1996) "Biological and cognitive foundations of intelligent data fusion." *IEEE Trans., Syst., Man, Cybern.*, vol.26, pp.42-51, January 1996.

Myers, B.A. (1996). "A brief history of Human Computer Interaction technology." Human Computer Interaction Institute, School of Computer Science, Carnegie Mellon University, Pittsburg, PA. http://reports-archive.adm.cs.edu/anon/1996/-CMU-CS-96-163.ps.

Mynatt, E.D. and Weber, G. (1994). "Nonvisual presentation of graphical interfaces: contrasting two approaches." CHI 94, pp.166-172.

Navon, D. & Gopher, D. (1979). "On the Economy of the Human Processing Systems," *Psychological Review, 86*, pp. 254-255.

Nicolelis, M. and Chapin, J.K. (2002). "Controlling robots with the mind." *Scientific American*, October.

Nilsson, L., Ohlsson, K., & Ronnberg, J. (1977). "Capacity Differences in Processing and Storage of Auditory and Visual Input," In S. Dornick (Ed.), *Attention and Performance VI*. Hillsdale, NJ: Erlbaum.

Oviatt, S., DeAngeli, S., and Kuhn, K. (1997) "Integration and synchronization of input modes during multimodal HCI," *Proceedings Conference Human Factors in Computing Systems (CHI '97)*, Atlanta, GA, pp.415-422.

Pashler, H. (1998). *The Psychology of Attention*. Cambridge, MA: MIT Press.

Pavlovic, V.I., Sharma, R., and Huang, T.S. (1997), "Visual Interpretation of hand Gestures for Human-Computer Interaction: A Review." *IEEE Transactions on Pattern Analysis and Machine intelligence*, Vol.19, No.7, July 1997.

Pavlovic, V.I., Berry, G.A., and Huang, T.S. (1997). "Integration of Audio/Visual Information for Use in Human-Computer Intelligent Interfaces." IEEE.

Raskin, J. (2000). *The Human Interface*. Boston: Addison Wesley.

Rasmussen, J. (1980). "The Human as a System's Component." In H.T. Smith & T.R. Green (Eds.), *Human interaction with computers*. London: Academic Press.

Rasmussen, J. (1986). *Information Processing and Human-Machine Interaction: An Approach to Cognitive Engineering*. New York: North Holland.

Rupert, A. (1998). "Haptics as the Most Intuitive Spatial Orientation System," *Proceedings of the Third Annual Symposium and Exhibition on Situational Awareness in the Tactical Air Environment*, Piney Point, MD (June).

Sanders, M.S. and McCormick, E.J. (1993). *Human Factors in Engineering and Design*. New York: McGraw-Hill, Inc.

Selcon, S., Taylor, R., and Shadrake, R. (1992). "Multi-Modal Cockpit Warnings: Pictures, Words, or Both?," *Proceedings of the Human Factors Society 36th Annual Meeting*, Atlanta, pp. 57-61.

Sellen, A., Kurtenbach, G., & Buxton, W. (1992). "The Prevention of Mode Errors Through Sensory Feedback," *Human Computer Interaction, 7*(2), pp. 141-164.

Sharma, R., Pavlovic, V.L., Huang, T.S. (1998) "Toward Multimodal Human-Computer Interface." *Proceedings of the IEEE*, vol.86, no.5.

Tan, D., Stefanucci, J., Proffitt, D., & Pausch, R. (2002). "Kinesthetic Cues Aid Spatial Memory," *Presented at the CHI 2002 Conference on Human Factors in Computing Systems*, Minneapolis, MN.

Tan, H., Lim, A., & Traylor R. (2000). "A psychophysical study of sensory saltation with an open response paradigm," In *Proceedings of the ASME Dynamic Systems and Control Division*, Vol. 69-2, pp. 1109-1115.

Teichner, W. & Krebs, M. (1974). "Laws of Visual Choice Reaction Time," *Journal of Experimental Psychology, 14*, pp. 1-35.

Vo, M. and Waibel, A. (1997). "Modeling and Interpreting Multimodal Inputs: A Semantic Integration Approach," Technical Report CMU-CS-97-192, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA.

Wickens, C. (1980). "The Structure of Attentional Resources," in R. Nickerson (Ed.), *Attention and Performance VIII* (pp. 239-257). Hillsdale, NJ: Erlbaum.

Wickens, C. (1984). "Processing Resources in Attention," In R. Parasuraman & R. Davies (Eds.), *Varieties of Attention* (pp. 63-101). New York: Academic Press.

Wickens. C. (1991). "Processing Resources and Attention," In D. Damos (Ed.). *Multiple Task Performance*. London: Taylor & Francis.

Wickens, C. & Hollands, J. (1999). *Engineering Psychology and Human Performance (3rd Edition)*. Upper Saddle River, NJ: Prentice-Hall.

Wickens, C. & Liu, Y. (1988). "Codes and Modalities in Multiple Resources: A Success and a Qualification." *Human Factors, 30*, pp. 599-616.

Wickens, C., Sandry, D., & Vidulich, M. (1983). "Compatibility and Resource Competition between modalities of input, central processing, and output," *Human Factors, 25(2)*, pp. 227-248.

Wickens, C., Vidulich, M., & Sandry-Garza, D. (1984). "Principles of S-C-R Compatibility with Spatial and Verbal Tasks: The Role of Display-Control Location and Voice-Interactive Display-Control Interfacing," *Human Factors, 26(5)*, pp. 533-543.

Woodworth, R.S. & Schlossbert, H. (1965). *Experimental Psychology.* New York: Holt, Rinehart, and Winston.

Youngblut, C., Johnston, R., Nash, S., Wienclaw, R., & Will, C. (1996). "Review of Virtual Environment Interface Technology," IDA Paper P-3186, Institute for Defense Analyses, Alexandria, VA.